Facial Expression Recognition in the Presence of Speech using Blind Lexical Compensation

Soroosh Mariooryad, *Student Member, IEEE,* and Carlos Busso, *Senior Member, IEEE,*

Abstract—During spontaneous conversations the articulation process as well as the internal emotional states influence the facial configurations. Inferring the conveyed emotions from the information presented in facial expressions requires decoupling the linguistic and affective messages in the face. Normalizing and compensating for the underlying lexical content have shown improvement in recognizing facial expressions. However, this requires the transcription and phoneme alignment information, which is not available in broad range of applications. This study uses the asymmetric bilinear factorization model to perform the decoupling of linguistic and affective information when they are not given. The emotion recognition evaluations on the IEMOCAP database show the capability of the proposed approach in separating these factors in facial expressions, yielding statistically significant performance improvements. The achieved improvement is similar to the case when the ground truth phonetic transcription is known. Similarly, experiments on the SEMAINE database using image-based features demonstrate the effectiveness of the proposed technique in practical scenarios.

Index Terms—facial expressions, emotion recognition, face analysis, factor analysis, bilinear model.

1 Introduction

The human face conveys an intricate blend of information including age, gender, ethnicity, identity, personality, intentions, and emotions. In addition, speech articulation greatly affects the facial appearance [1], [2]. All these aspects regulate human communication at various levels. While we can process this complex blend of information to effectively separate the underlying factors, the presence of all these sources of variability makes this process a challenging task for machine-based inferences. For instance, the performance of conventional face recognition systems drops when the facial expressions deviate from neutral emotion [3]–[5]. Similarly, effective emotion recognition systems should be robust against inherent variations in the facial structure across different individuals [6].

In the context of facial expression recognition, previous studies have proposed methods to compensate for head pose [7] and subjects differences [6]. However, it is not clear how to reduce the variability introduced by the spoken message during spontaneous interactions. This is an important challenge in the development of robust emotion recognitions systems, since speech articulation affects the orofacial area including lips and jaw [8]. The same facial region plays a key role in conveying emotions [9]. Notice that if the underlying lexical information is not considered, neutral facial poses may be confused with emotional expressions. For example, uttering the phoneme /ey/ may be confused with a smile, and uttering the phoneme /o/ may be confused with surprise [10]. Our previous studies reveal the interplay between verbal and affective messages in facial expressions [1], [2], [11]. Therefore, to extract reliable emotional information, facial expressions should be decoupled into the underlying factors.

We have shown that the lower facial area is highly dependent on the articulated message, which masks the emotional information conveyed through this area [2]. Our analysis and emotion recognition evaluations indicated that making use of the phonetic transcription to compensate for the articulated message can significantly improve the emotion discrimination of facial features extracted over the orofacial area [2], increasing emotion recognition rates [12]. In those studies, we assumed that the phoneme alignment of the spoken message was available for the lexical compensation approach. Hence, we refer to that approach as supervised lexical compensation (supervised in the sense that the transcription of lexical content is available). Assuming the availability of the phonetic alignment might not be realistic for real-world applications. Although *automatic speech recognition* (ASR) systems can provide this information, its use introduces an extra source of complexity to the system. This paper proposes a blind lexical compensation approach to reduce the variability due to speech articulation in facial expressions for robust emotion recognition. The approach does not assume that the lexical information is available.

This paper proposes to factorize facial movements into the underlying phonetic and emotional content using the asymmetric bilinear model [13], [14]. This factor separation model was originally developed to separate style from content in machine learning applications. In our problem, the content represents the articulatory configurations for each phoneme/viseme, and the style

• *This study was funded by Samsung Telecommunications America and* National Science Foundation *(NSF) grants IIS-1217104 and IIS-1329659.*
The authors are with the Multimodal Signal Processing (MSP) laboratory, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: soroosh.ooryad@utdallas.edu, busso@utdallas.edu).

represents the emotional information. During the training step, the phoneme and emotional labels are provided to capture the interaction between these two factors. However, in the testing step, neither of these labels are available. They are estimated with an *expectation maximization* (EM) approach that iterates to obtain the best fit for the testing data. It implicitly predicts the phonemes, which are used to compensate for the lexical variability. We evaluate this approach, which is referred to as blind lexical compensation, with the IEMOCAP database achieving performances similar to the case where the ground truth phonetic labels are used to compensate for the lexical variability (supervised lexical compensation approach presented in [12]). We also present experiments on the SEMAINE database showing emotion recognition improvements with the proposed blind lexical compensation method.

The rest of the paper is organized as follows. Section 2 motivates the problem and presents related studies. This section also summarizes our previous studies on factor analysis of facial features. Section 3 describes the IEMOCAP and SEMAINE databases. Section 4 describes the proposed blind lexical compensation method. Section 5 presents extensive emotion recognition experiments which show the benefits of using the proposed blind lexical compensation. Section 6 summarizes the findings of the paper and discusses the implications in emotion recognition systems.

2 Background

2.1 Related Work

Despite the advancements in identifying prototypical expressions in static images, modeling the dynamic aspects of facial expressions is still an ongoing research problem. Studying the significance of temporal information, in contrast to posed expressions, and its relation to speech articulation is among the key open aspects identified by pioneers of the field [15]. Perceptual evaluation has been used to understand the role of dynamic information in facial expression. While dynamic information may not provide complementary information in full-blown expressions [16], its role in subtle and natural emotions is clearly important in perceiving emotions [17]. Temporal dynamics of facial expressions also play a key role in characterizing other cognitive states such as embarrassment, amusement, pain and mood [18]. Therefore, dynamic processing of facial expressions has recently gained more interest [19]–[21]. In spite of these efforts, the effect of articulation on the dynamics of facial expressions during spontaneous interaction is often neglected.

Studies have shown that facial features provide better discrimination on the valence domain (positive versus negative) [22], and speech features provides better discrimination on the arousal domain (calm versus active) [23]. These observations suggest that fusing these complementary modalities results in higher accuracy. However, Picard [24] mentioned that fusing these two modalities is not an easy task due to the differences in mouth movements between speaking and non-speaking conditions. Shah et al. [25] compared the performance of facial emotion recognition under both conditions. The performance of their classifier drops when the subjects were speaking. To reduce the mismatch, they proposed to train their system with data from speaking recordings. Despite the increase in performance, the classifiers achieved lower accuracy than the ones achieved during the non-speaking condition. They also concluded that the movements in the orofacial area misled the classifiers in detecting anger and disgust, compared to the case when the classifiers were trained with non-speaking faces. Due to the high lexical variability in the lower facial area, some studies considered only features derived from the middle and upper face area, ignoring all the emotional information conveyed in the orofacial area [26]. Chen and Huang [27] claimed that articulation affects not only the lower facial area, but also the upper part of the face, since eyebrows convey other linguistic messages such as emphasis. Therefore, their proposed multimodal emotion recognition system relied on speech only when the subject was speaking and on facial expression only during the silent segments. Likewise, Yong et al. [28] only used the non-speaking segments to analyze the emotions in the face.

Although some approaches neglect the emotion-discriminative information conveyed through the lower part of the face, previous studies suggest the presence of affective message in the orofacial area [26], [29], [30]. The orofacial appearance is closely tied with the acoustic properties of the speech signal [8]. Therefore, expressive behaviors modulate the orofacial area during spontaneous interactions. Our previous analyses demonstrated the interplay between verbal and affective messages unfolded in localized facial areas, especially the orofacial area [2], [11]. Subjective studies showed that different emotions can be conveyed through different facial areas [31], [32]. For instance, Hoffmann et al. [32] demonstrated that emotions such as happiness and sadness are better perceived from the lower part of the face. These studies suggest that the orofacial area conveys important emotion-discriminative information that a facial expression recognition system should capitalize on to produce robust classification.

An important step to correctly exploit and model the emotional information in the orofacial area is to compensate for the lexical variability due to the articulation process [10]. Zeng et al. [33], [34] used a smoothing technique to attenuate the effect of speech in facial expressions. The approach consisted in averaging the facial features over 10 consecutive frames (i.e., 333ms). Wu et al. [35] proposed an eigenface-based method to convert speaking faces into non-speaking faces to compensate for the articulation effect in the mouth area.

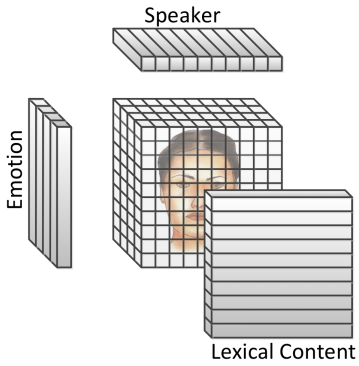


Fig. 1. Speaker, lexical content and emotions are the three factors modulating the variability of facial features.

2.2 Relation to Our Prior Work

In Mariooryad and Busso [2], we proposed to study facial expression as the result of three main factors that modulate the variability of facial features: speaker, lexical and emotional contents. Figure 1 illustrates the concepts of facial behaviors resulting from the interplay between these three factors. Based on this assumption, we proposed a factor analysis framework to identify the dependency of the facial areas on these factors. The analysis relies on the IEMOCAP database (see Sec. 3.1). Our findings clearly demonstrate the dominant influence of speech articulation on the orofacial features. Figure 2 depicts the dependency of each of the factors ($F \in \{Speaker, Lexical, Emotion\}$) and each facial point along the three directions (X – left/right; Y – up/down; Z – in/out). Darker colors indicate higher dependency. This analysis shows that speaker dependency is almost uniformly distributed across the entire face and it is the weakest factor among all three. The orofacial area is completely dominated by the lexical variability. The effect of emotion is mainly evident in the middle and upper face regions. Our analysis also showed that constraining the models on the lexical content makes the emotion variability the dominant factor in the mouth area too [2]. This data-driven analysis unveils the interplay between emotional and lexical content on the orofacial area. It also shows that compensating for the underlying lexical variability can uncover the conveyed emotional message. This study takes up that challenge of exploring algorithms to reduce the lexical dependency on facial features around the orofacial area. The contribution of this study is a blind lexical normalization scheme that does not requires the underlying transcription to compensate for the lexical information, providing an appealing solution for practical applications.

3 Databases

This study uses the *interactive emotional dyadic motion capture* (IEMOCAP) [36] and the *sustained emotionally coloured machine-human interaction using nonverbal expressions* (SEMAINE) databases [37]. The IEMOCAP database contains motion capture data, which is used to perform the facial factor analysis (Fig. 2) and also emotion recognition experiments (Sec. 5.1). To show the potential of the proposed blind lexical compensation method in real-world applications, we perform emotion recognition experiments on image-based features extracted from the SEMAINE database (Sec. 5.2). This section describes these two corpora.

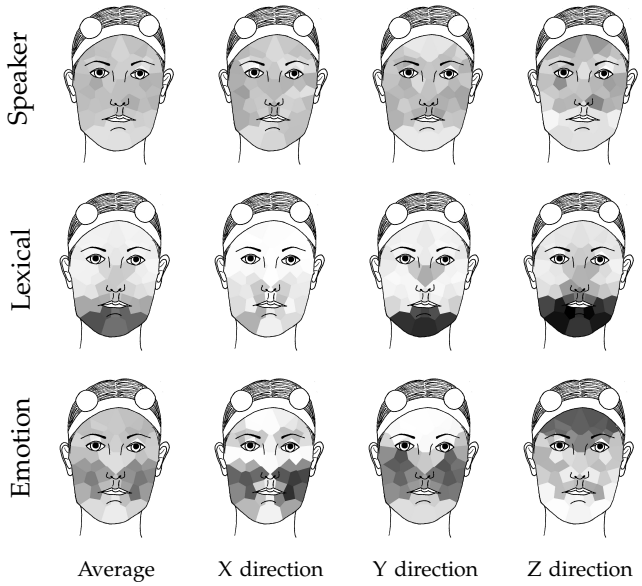


Fig. 2. The dependency of the factors (speaker, lexical and emotion) and different areas on the face according to the factor analysis technique introduced in [2]. Darker color represent higher dependency.

(a) Setup (b) Layout (c) Selected markers

Fig. 3. (a) Recording setup for the IEMOCAP database, (b) layout of the motion capture markers in the IEMOCAP database, and (c) selected markers used in our previous analysis [2], which are used for emotion recognition experiments (Sec. 5).

3.1 The IEMOCAP Database

The IEMOCAP corpus is an audiovisual database to study expressive human interactions recorded with 10 speakers (five male and five female). In five sessions, an actor and an actress performed scripted plays, and improvised hypothetical situations. The scenarios and situations are deliberately chosen to evoke emotional reactions (e.g., losing baggage at airport and getting school admission). The recordings include speech, video and

TABLE 1
Number of speaking turns for the selected emotions in the IEMOCAP database.

Emotions	Happiness	Anger	Sadness	Neutral	All
number	838	574	630	585	2627

motion capture data. The motion capture data is sampled at 120 Hz. Only one of the actors wore the motion capture markers at a time due to the limitations in motion capture recordings (see Fig. 3(a)). Therefore, each session is recorded twice to collect facial expressions from both actors. To capture precise facial movements 53 markers are placed on the face following the layout described in Figure 3(b). After compensating for head translation and rotation, the detailed facial expression movements are extracted. Given the inherent variations in the facial structures across speakers in the corpus, we normalize the markers using speaker-dependent z-normalization, where we estimate the mean and standard deviation using all the recordings from each speaker.

The conversations are segmented into speaking turns. The turns are manually transcribed. The phoneme and word boundaries are obtained with forced-alignment algorithm. The turns are emotionally annotated by three evaluators in terms of discrete categorical labels using the following labels: anger, happiness, sadness, surprise, disgust, frustration, fear, excitement, neutral state and other. Due to the overlap between happiness and excitement, and following the conventional approach on this database [12], [38], we have merged these two categories. We select the samples for which the evaluators reached agreement using majority of votes as a criterion. We selected the four most frequent emotions in the database for our experiments (i.e., happiness, anger, sadness and neutral state). Table 1 summarizes the number of samples per emotion. These speaking turns correspond to the segments from the actors while wearing the markers. The database and related preprocessing steps are described in Busso et al. [36].

3.2 The SEMAINE Database

SEMAINE is another audiovisual database containing recordings of users interacting with an operator. The operator plays the role of *sensitive artificial listener* (SAL) agents [39] with different personalities (happy, gloomy, angry and pragmatic) to evoke emotional reactions from the users. The sessions are segmented into speaking turns, which are manually transcribed. We also use forced-alignment to extract the underlying phoneme boundaries. Multiple annotators (from 2 to 8) have emotionally annotated the users' videos in 52 interaction sessions. The annotations include the degree of arousal (i.e., calm versus active), valence (i.e., negative versus positive) and other continuous attributes describing affective states. The FEELTRACE toolkit [40] is used to continuously annotate these emotional descriptors in the interval [-1, +1]. We have shown in our previous work that this annotation scheme, which involves watching the video, perceiving the affective content and responding by moving mouse cursor, introduces an intrinsic delay between the behaviors and the annotations [41]. We have employed a maximum mutual information criterion to estimate that delay [41], and this study uses these values to correct the annotations. The ratings are averaged across the annotators to obtain a single temporal trace for each session. Then, the average value along the duration of each utterance is assigned to speaking turns as the emotion descriptor. Similar to our previous studies [41], [42], the values are clustered into two (e.g., low versus high arousal) and three (e.g., low, medium and high valence) classes using the K-means algorithm (with $K = 2, 3$) to obtain discrete labels per speaking turn.

TABLE 2
Set of *action units* (AUs) extracted from lower part of the face using CERT [44].

AU	description	AU	description
AU 6	Cheek Raise	AU 20	Lip stretch
AU 10	Lip Raise	AU 23	Lip Tightener
AU 12	Lip Corner Pull	AU 24	Lip Presser
AU 14	Dimpler	AU 25	Lips Part
AU 15	Lip Corner Depressor	AU 26	Jaw Drop
AU 17	Chin Raise	AU 28	Lips Suck
AU 18	Lip Pucker		

The proposed approach is evaluated in Sec. 5.2 using facial features extracted from videos. We rely on *action units* (AUs) as high level representations of facial expressions. The AUs are defined in the *facial action coding system* (FACS) [43], and represent the degree of contraction or relaxation of one or multiple facial muscles. We extract the AUs using the *computer expression recognition toolbox* (CERT) [44]. This toolkit gives frame-by-frame AU values based on models trained with Gabor-based features. Since the focus of this study is on the speech articulation effect on the orofacial area, we only choose AUs corresponding the lower facial areas (see Table 2). The CERT package failed to detect the user's face in 8 out of 52 sessions that were emotionally annotated (sessions 82, 88-91, 95-97). Thus, the study only considers the remaining 44 sessions recorded from nine unique users. After removing the turns for which the forced-alignment approach failed to detect the phoneme boundaries, 939 speaking turns are selected for the experiments. We also evaluate the performance of the proposed solution during segments where the "user" is listening to the "operator" (non-speaking turns – see Sec. 5.2.1). We use 814 of these segments.

4 Blind Lexical Compensation

In our previous work, we proposed supervised lexical compensation methods by using phoneme alignments using transcriptions, which is an important limitation

TABLE 3
The mapping from phoneme to viseme proposed by Lucey et al. [45]. The frequency of occurrence of the phonemes in the selected portion of the IEMOCAP database is given in column #.

#	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

[12] (details are given in Sec. 5.1.4). We aim to build a blind lexical compensation method that factorizes the contribution of emotional and lexical content that does not require transcriptions.

Previous studies have shown the influence of lexical content on the orofacial area. Furthermore, the study in Mariooryad and Busso [2] demonstrated that compensating for the lexical content is useful only on the orofacial area. Therefore, we explore a blind lexical normalization scheme tailored to facial features derived from the lower part of the face (i.e., the facial area with the strongest lexical dependency). Notice that features from other facial areas are less affected by the lexical information, so a simpler lexical-independent model can be used. The rest of the study considers only the 15 markers around the mouth area, highlighted in the shaded area of Figure 3(c). These markers are identified by our study on factor analysis on facial features [2].

We employ phonemes as our lexical unit to retain adequate number of samples for building lexical-dependent models. Phonemes have shorter duration than words or syllables, and are more affected by coarticulation effects. However, the compacted number of phonetical classes to describe lexical content facilitates the training of robust lexical-dependent models. Since certain phonemes produce similar facial appearance (e.g., phonemes /F/ and /V/), we consider *visemes* – visual phonemes that cluster acoustic phonetic units with similar orofacial configuration. We apply the phoneme-to-viseme mapping given by Lucey et al. [45] to represent the lexical content (Table 3). In total 13 non-silence visemes are used in the experiments. Table 3 reports the number of occurrences of the phonemes in the database.

4.1 Overview of the Asymmetric Bilinear Model

The blind lexical compensation approach builds upon the asymmetric bilinear model introduced by Tenenbaum and Freeman [13], [14]. The original model was proposed to separate style from content (e.g., in isolated character recognition *content* represents different letters and *style* represents different fonts). The model assumes N_s different styles, and N_c different contents. Consider the following equations,

$$\mathbf{y}^{sc} = A^s \mathbf{b}^c = \sum_{i=1}^{N_c} b_i^c \mathbf{a}_i^s \quad (1)$$

$$A^s = [\mathbf{a}_1^s, \mathbf{a}_2^s, \ldots, \mathbf{a}_{N_c}^s] \quad (2)$$

where $\mathbf{y}^{sc}$ is a $K \times 1$ observation vector conveying content c in style s. The asymmetric bilinear model assumes that $\mathbf{y}^{sc}$ lies close to a lower dimensional J-subspace ($J < K$) spanned by the columns of the $K \times J$ matrix A^s, associated with style s. Every instance in style s is approximated by a linear combination of the column vectors ($\mathbf{a}_i^s$) of the style-dependent matrix A^s. The $J \times 1$ vector $\mathbf{b}^c$ represents the underlying content c, containing the combination coefficients across all styles. The content vector $\mathbf{b}^c$ is independent of the style. Therefore, it is possible to reproduce content c in a different style, by replacing the matrix A^s from one style (i.e., s_1) to another style (i.e., s_2) (see Eq. 1). The value for J is generally small (in our study it varies from 2 to 4), providing a low dimension subspace for the column space of A^s, which represents the direction with more variability for the style of s. We can use the independence between the style (A^s) and content ($\mathbf{b}^c$) in the bilinear model to decouple these factors.

The bilinear model offers an ideal framework to separate the contribution of lexical (i.e., "content") and emotional (i.e., "style") information. We represent the content with visemes ($N_c = 13$) and the style with the underlying emotional state ($N_s = 4$). The column vectors of A^s form a subspace for a specific style s (emotion) to reconstruct the content (visemes) in that style. The bilinear structure of the models forces the basis vector $\mathbf{a}_i^s$ to have similar role across styles.

Figure 4 visualizes the implementation of the bilinear model applied to facial features. The figure provides the mean (solid line) and standard deviation (dashed line) of the 15 motion capture markers for the orofacial area. We generate these figures by training the bilinear models for visemes /aa/ ($\mathbf{b}^{/aa/}$) and /ch/ ($\mathbf{b}^{/ch/}$), for the emotional classes neutral (A^{Neu}), happiness (A^{Hap}), anger (A^{Ang}), and sadness (A^{Sad}). Section 4.2 gives the details on the training procedure. Vectors $\mathbf{b}^{/aa/}$ and $\mathbf{b}^{/ch/}$ are the viseme-dependent vectors. Applying these vectors to each of the emotion-dependent subspaces gives that

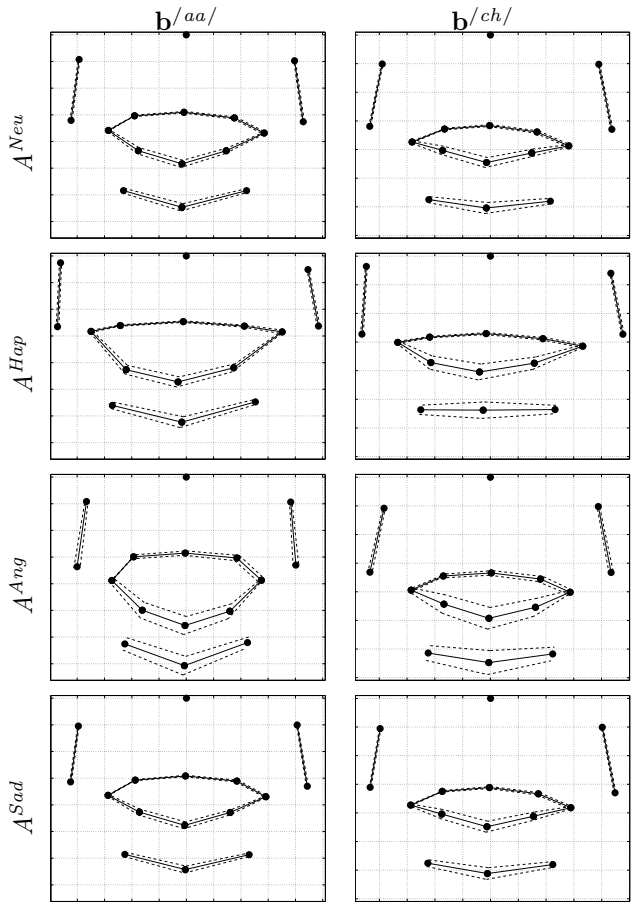


Fig. 4. Illustration of factor separation with the asymmetric bilinear model on the 15 orofacial motion capture markers identified in Figure 3(c). The mean and standard deviation of the markers are depicted with the solid and dashed lines, respectively. Multiplying the viseme-dependent vector ($\mathbf{b}^{/aa/}$ and $\mathbf{b}^{/ch/}$) to each of the emotion-dependent basis matrices (A^{Neu}, A^{Hap}, A^{Ang} and A^{Sad}) gives the average value of the features for that pair of viseme and emotion.

viseme in a specific emotion. The figure shows that multiplying the emotion-dependent matrix ("style") to the vector representing a given viseme ("content") produces different visual appearance. For example, the opening of the mouth increases for happiness and anger. However, the overall lip configuration associated to the underlying viseme is preserved, regardless of the emotional content.

4.2 Bilinear Model Training

The goal of the training step in the bilinear model is to estimate the content vector $\mathbf{b}^c$ for each viseme c, and the style matrix A^s for each emotion s (see Fig. 6). For training, we assume that the emotional labels, and the phonetic alignment of the underlying visemes are available. When the number of samples with a given style and content are similar (balanced case), Tenenbaum and Freeman [13] gave a closed-form solution based on *singular value decomposition* (SVD). The approach estimates the average observation vector $\bar{\mathbf{y}}^{sc}$, for a given style s and content c. Then, these vectors are stacked to create a matrix $\bar{Y}$,

$$\bar{Y} = \begin{bmatrix} \bar{\mathbf{y}}^{11} & \cdots & \bar{\mathbf{y}}^{1N_C} \\ \vdots & \ddots & \vdots \\ \bar{\mathbf{y}}^{N_S 1} & \cdots & \bar{\mathbf{y}}^{N_S N_C} \end{bmatrix} = USV^T \tag{3}$$

where matrices U, S and V are computed with SVD. The first J columns of matrix US and the first J rows of V^T define the style matrices A^s, and content vectors $\mathbf{b}^c$, respectively (see details in Tenenbaum and Freeman [13]). For the case when the classes are not balanced, as in our problem, there is an iterative approach that minimizes the reconstruction error of Equation 1 [14]. The approach initializes the values for A^s and $\mathbf{b}^c$ using the SVD-based closed-form solution. Then, it estimates A^s using Equation 4, assuming that $\mathbf{b}^c$ is fixed. The vector $\mathbf{m}^{sc}$ is the sum of the observations in style s and content c (see Eq. 6). The variable n^{sc} is the number of samples in the training set with style s and content c. Then, $\mathbf{b}^c$ is estimated with Equation 5, assuming that A^s is fixed. We iterate the algorithm 10 times in our implementation.

$$A^s = \left[\sum_c \mathbf{m}^{sc}\mathbf{b}^{c^T}\right]\left[\sum_c n_{sc}\mathbf{b}^c\mathbf{b}^{c^T}\right]^{-1} \tag{4}$$

$$\mathbf{b}^c = \left[\sum_s n^{sc}A^{s^T}A^s\right]^{-1}\left[\sum_s A^{s^T}\mathbf{m}^{sc}\right] \tag{5}$$

$$\mathbf{m}^{sc} = \sum_{s,c}\mathbf{y}^{sc} = n^{sc}\bar{\mathbf{y}}^{sc} \tag{6}$$

4.3 Bilinear Model Testing

This section describes the algorithms for testing the bilinear model. The testing step assumes that an utterance is given with a set of unlabeled visemes in an unknown emotion, which is not necessarily one of the trained emotions. From the training step (Sec. 4.2), we have the content vector $\mathbf{b}^c$ for each viseme class c, and the style matrices A^s for each emotion class s. The task in the testing step is to find the best style matrix A^{s^*} for an utterance without any lexical information – the underlying visemes are unobserved. However, we assume that the segmentation of the visemes is known (we relax this constraint in Sec. 5.1.3).

Since the content (viseme) labels, c, are unknown, we cannot use Equation 4 to compute matrix A^{s^*}. Instead, Tenenbaum and Freeman [14] proposed to embed the bilinear model in a Gaussian mixture distribution referred to as *separable mixture model* (SMM) and run the *expectation maximization* (EM) algorithm to estimate the style subspace A^{s^*} associated with the testing utterance. For the training procedure (Sec. 4.2), a segment was assigned

to a given content (viseme). The key idea of the EM procedure is replacing this hard assignment with a soft assignment. Let $p(s^*, c|\mathbf{y})$ be the probability that sample $\mathbf{y}$ belongs to content/viseme c and style/emotion s^*. Notice that the assignment of style is not important for the EM procedure, since the approach assumes that the testing utterance may belong to a different style (emotion) not included in the training. The testing procedure starts by initializing A^{s^*} with the average value across the trained styles ($A^{s^*} = (A^{Hap} + A^{Ang} + A^{Sad} + A^{Neu})/4$). The E-step consists in estimating the soft assignment $p(s^*, c|\mathbf{y})$ using Equation 7. The approach assumes that the probability of an observation $\mathbf{y}$ being generated by content c and style s^* follows a multivariable Gaussian distribution with mean vector $A^{s^*}b^c$ and covariance matrix $\Sigma = \sigma^2 I$ (see Eq. 8). A second assumption is that $p(s^*, c)$, the prior probability of content c and style s^*, is uniformly distributed (this probability can also be estimated from the data). Under these assumptions, we estimate $p(s^*, c|\mathbf{y})$ with Equation 7.

$$p(s^*, c|\mathbf{y}) = \frac{p(\mathbf{y}|s^*, c)p(s^*, c)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|s^*, c)p(s^*, c)}{\sum_c p(\mathbf{y}|s^*, c)p(s^*, c)} \quad (7)$$

$$p(\mathbf{y}|s^*, c) \sim \mathcal{N}(A^{s^*}\mathbf{b}^c, \sigma^2 I) \quad (8)$$

The M-step uses the soft assignment $p(s^*, c|\mathbf{y})$ to estimate A^{s^*}. First, we estimate the number of samples associated with content c and style s^* (Eq. 9). Then, we estimate $\hat{\mathbf{m}}^{s^*c}$, the sum of the observation vectors for c and s^* (Eq. 10). We estimate A^{s^*} with these variables (Eq. 11).

$$\hat{n}^{s^*c} = \sum_y p(s^*, c|\mathbf{y}) \quad (9)$$

$$\hat{\mathbf{m}}^{s^*c} = \sum_y p(s^*, c|\mathbf{y})\mathbf{y} \quad (10)$$

$$A^{s^*} = \left[\sum_c \hat{\mathbf{m}}^{s^*c}\mathbf{b}^{c^T}\right]\left[\sum_c \hat{n}_{s^*c}\mathbf{b}^c\mathbf{b}^{c^T}\right]^{-1} \quad (11)$$

We use ten iterations as the stopping criterion. This model has two parameters: J, the number of columns of the style matrices A^s; and σ, the standard deviation associated with the Gaussian distribution for $p(\mathbf{y}|s^*, c)$ (see Eq. 8). Similar to Tenenbaum and Freeman [13], these parameters are set using cross-validation on the training set (see Sec. 5). We run the EM algorithm with unlabeled viseme(s) in an unknown emotion to find the best viseme label(s), captured by $p(s^*, c|\mathbf{y})$, and the best style subspace, captured by A^{s^*}. Notice that the extraction of A^{s^*} uses the vectors $\mathbf{b}^c$ extracted from the bilinear training step. This approach implicitly compensates for the lexical variability.

There are various frameworks to separate multiple factors affecting a given variable including *nuisance attribute projection* (NAP) [46], *linear discriminant analysis* (LDA) [47], *probabilistic linear discriminant analysis* (PLDA) [48], *joint factor analysis* (JFA) [49] and general n-order tensor-based factor analysis [50]. The key advantage of using the asynchronous bilinear model is that multiple instances (visemes in this work) can be simultaneously used by constraining them to have the same style (emotion), adapting the style subspace. This approach results in less degrees of freedom for the adapted style and, as a result, it gives more reliable estimation of the style factor (i.e., emotion). While the symmetric bilinear model can also be used to address this problem, we rely on the asymmetric model since it provides more flexibility in terms of style space adaptation, specially when the testing data does not exactly match the pre-trained style models [14].

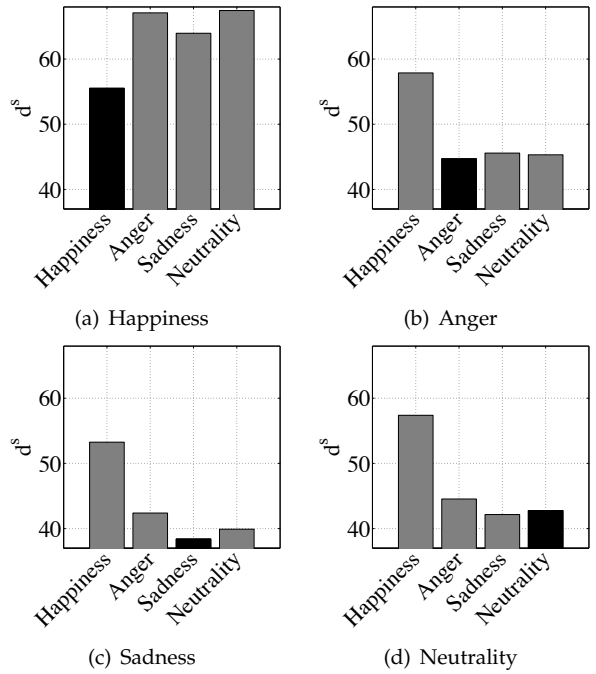
4.4 Proposed Features For Emotion Classification

The bilinear model was proposed for solving classification (i.e., recognizing the underlying content in unknown styles), extrapolation (i.e., generating content in a new style) or translation (i.e., changing a new style into a known style) problems [13]. We make use of the testing procedure designed for the classification problem, which consists in recognizing the content based on $p(c|\mathbf{y})$ after estimating A^{s^*} in the EM iterations. However, there is an important distinction between how the bilinear model has been used in other studies and our proposed method. Here, we are not interested in classifying the visemes, but the style (emotion) based on the adapted subspace (unknown style) achieved after testing the bilinear model. We use the bilinear model to extract emotional discriminant features at utterance level, which are robust against lexical variations (A^{s^*}). As discussed in Section 4.2, the adapted subspace A^{s^*} is independent of the underlying visemes and we use its entries as our features.

If the test utterance conveys the emotion $s^e \in \{Hap, Ang, Sad, Neu\}$, we expect to observe higher similarity between A^{s^*} and A^{s^e} than with any other matrix with a different style (i.e., A^{s^*} will converge closer to A^{s^e}). To validate this assumption, we use the Frobenius norm of the difference between pair-wise elements of the subspace matrices as a measure of similarity (Eq. 12).

$$d^s = \left\| A^{s^*} - A^s \right\|_F \quad (12)$$

For each testing emotion, Figure 5 depicts the average d^s values for each of the subspaces obtained from the training data (the lower the value, the closer the matrices). For instance, Figure 5(a) indicates that A^{Hap} obtained from the training data is the closest subspace to A^{s^*} when it is adapted with happy testing samples. The emotions anger and sadness follow similar patterns – their subspaces converge closer to the subspace of their corresponding emotion. Although the samples with neutral emotion produce a style matrix closer to the one for sadness, the matrix A^{s^*} is still close to A^{Neu}.



(a) Happiness (b) Anger

(c) Sadness (d) Neutrality

Fig. 5. The average Frobenius norm of difference between adapted subspace and four trained subspace. The figures are depicted for each testing emotion independently.

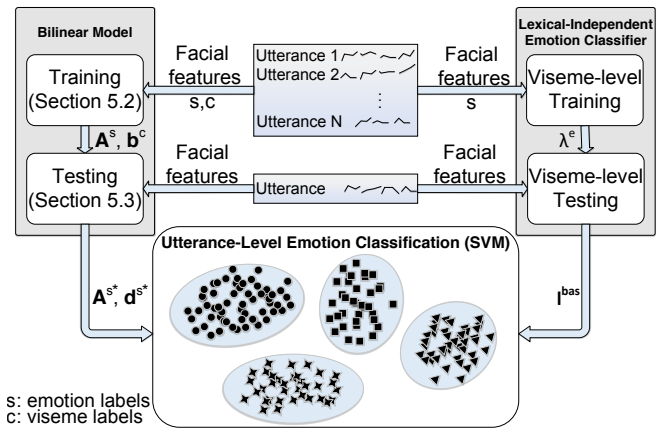


Fig. 6. The proposed blind lexical compensation scheme for robust emotion classification.

Note that neutral emotion is the most difficult state to characterize, and it is often confused with other emotions [42], [51]–[53]. Figure 5 shows the effectiveness of the proposed approach in compensating for the lexical variability, without transcriptions, to find the best underlying emotion. The figure also suggests that the subspace for happiness is significantly different from the subspaces of other emotions.

Figure 6 depicts the proposed two-layer training procedure for our emotion classifier. Since the utterance level classifier does not directly use the facial features, we refer to inputs of this classifier as meta-features. The first layer consists of training the viseme-based bilinear model (Sec. 4.2). The second layer consists of training the utterance-level emotion classifier with the meta-features derived by presenting the utterances to the bilinear model (Sec. 4.3). In addition to A^{s^*}, we use two additional meta-feature sets. The first set is a 4-D vector $\mathbf{d}^{s^*} = [d^{Hap}, d^{Ang}, d^{Sad}, d^{Neu}]$ (see Eq. 12). We use this vector as a compact representation of A^{s^*}. The second feature set (Fig. 6) does not rely on the bilinear model. We separately train a lexical-independent viseme-level emotion classifier with facial features as the first layer of training (λ^e in Fig. 6). The viseme-level classifier provides the likelihoods for every viseme in the utterance. Notice that the emotional labels in the IEMOCAP database are annotated at the utterance level. During training, we assign the label of the utterance to each of its individual visemes. The results of this classifier is the emotional class likelihoods for each viseme in the utterance. Then, we use the mean, minimum and maximum of the four class likelihoods across the visemes of the utterance, creating a 12-D feature vector $\mathbf{l}^{bas}$ (class likelihoods for 4 emotions $\times$ 3 statistics). This vector $\mathbf{l}^{bas}$ is used as features in the second layer of training by the utterance-level emotion classifier (see Fig. 6). Extraction of $\mathbf{l}^{bas}$ does not make any distinction about the underlying visemes. Therefore, it can be considered as a baseline with no lexical compensation.

To evaluate the emotion of a testing utterance, the features are processed at the viseme level by the bilinear model to obtain A^{s^*} and $\mathbf{d}^{s^*}$, and by the lexical-independent emotion classifier to obtain the viseme-level class likelihoods ($\mathbf{l}^{bas}$). These features are used to recognize the emotions. Notice that the testing procedure assumes that the emotion and viseme labels are both unknown, but the viseme boundaries are given. However, our experiments in Section 5.1.3 shows that the blind lexical compensation approach can retain comparable performance even without the viseme boundaries.

The bilinear model processes all the visemes of the utterance at once to estimate A^{s^*} and $\mathbf{d}^{s^*}$. Notice that only some of the visemes across an utterance may be emotionally colored (e.g., we may observe "neutral" visemes in an angry utterance). This subset of "emotional" visemes in an utterance will affect the trained subspace for that emotion. During the testing step, they will also affect the adapted subspace providing emotion-discriminative information. This is one of the strength of the proposed approach, which is, up to some extent, robust against emotion variations across the visemes of an utterance.

5 Emotion Recognition Experiments

We present extensive set of emotion recognition experiments on the IEMOCAP database, where features are derived from motion capture data (Sec. 5.1). The effectiveness of the proposed method is also validated with experiments on the SEMAINE database, where the facial features are automatically extracted from the videos (Sec. 5.2).

TABLE 4
Optimized values of J and σ per cross-validation fold.

fold	1	2	3	4	5	6	7	8	9	10
J	2	3	2	2	3	2	4	2	2	2
σ	2	2	2	2	1	0.5	2	2	0.5	1

The evaluation uses linear kernel SVM, with *sequential minimal optimization* (SMO), both for the viseme-level (λ^e) and the utterance-level classifiers. We use the implementation provided by the WEKA toolkit [54]. We report the classification performance in terms of accuracy and macro-average F-score. We estimate F-score as follows. First, we compute the precision and recall rates for each of the emotional classes. Then, we estimate the average precision ($\bar{p}$) and recall ($\bar{r}$) rates across emotional classes. Then, we estimate the F-Score with Equation 13. This metric is not biased by unbalanced classification problems (classification performance at change is 25% for a four-class problem). We use the proportion hypothesis test to evaluate whether the differences in performance are statistically significant among different settings.

$$F - score = 2\frac{\bar{p} \cdot \bar{r}}{\bar{p} + \bar{r}} \tag{13}$$

5.1 Evaluations on the IEMOCAP Database

The evaluations on the IEMOCAP database follow a 10-fold *leave-one-speaker-out* (LOSO) cross-validation approach. The facial features are extracted from the motion capture data, which provides detailed facial information that is particularly useful to analyze the proposed approach in controlled recordings. As described in Section 4, we consider only the 15 markers around the mouth area for the experiments (see Fig. 3(c)). For a given viseme, we extract the following statistics of the trajectories of the markers: minimum, maximum, mean, standard deviation, median, lower quartile and upper quartile. These statistics are independently extracted from each marker, and for each direction. This gives a 315-D feature vector (15 markers $\times$ 3 directions $\times$ 7 statistics). This feature vector is consistently used in this section for the bilinear model and the viseme-level lexical-independent classifier. In each fold, data from nine speakers are used for training.

5.1.1 Performance of Blind Lexical Compensation

The classification evaluation for the proposed approach is implemented with the 10-fold LOSO cross-validation approach. In each fold, data from nine speakers are used for training. The data from the remaining speaker is only used for testing. In each of the ten folds, we further split the training data into two speaker-independent partitions. Partition A has data from four speakers, and it is used for training the bilinear model (Sec. 4.2). Notice that we train the bilinear model using the viseme and emotion labels. We also train the lexical-independent viseme-level classifier with this partition (λ^e in Fig. 6).

TABLE 5
Emotion recognition results of the proposed blind lexical compensation method. The evaluation includes correct and random viseme boundaries – IEMOCAP.

Feature Sets	With Boundaries Accuracy [%]	With Boundaries F-Score [%]	Random Boundaries Accuracy [%]	Random Boundaries F-Score [%]
$\mathbf{l}^{bas}$ [baseline]	53.81	52.18	54.09	51.19
$\mathbf{d}^{s^*}$	43.62	41.36	50.02	45.06
A^{s^*}	58.52	57.29	58.32	56.73
$\mathbf{l}^{bas}$, $\mathbf{d}^{s^*}$	54.58	52.92	55.54	52.35
$\mathbf{l}^{bas}$, A^{s^*}	60.36	59.10	58.81	57.36
$\mathbf{d}^{s^*}$, A^{s^*}	58.83	57.54	58.17	56.66
$\mathbf{l}^{bas}$, $\mathbf{d}^{s^*}$, A^{s^*}	60.25	58.94	58.17	56.80

Partition B has data from five speakers. Utterances of this partition are segmented into visemes and they are presented to the bilinear model (Sec. 4.3) to obtain A^{s^*} and $\mathbf{d}^{s^*}$ and also to the viseme-level classifier to obtain $\mathbf{l}^{bas}$. These meta-features are then used to train the utterance-level classifier (second layer of the proposed approach – see Sec. 4.4). We emphasize that partitions A and B are from the train set. The test set is only used to evaluate the classification performance. This approach ensures the generalization of the approach for new speakers.

We separately set the parameters J and σ of the model in each fold. For this purpose, we run a second cross-validation approach over partition B (data from five speakers used to train the emotion classifier). The dimension of the subspace of the style matrix J is set between 1 to 10. We evaluate the following values for $\sigma \in \{0.125, 0.25, 0.5, 1, 2\}$. This approach generates different sets of parameters per fold. Table 4 reports the optimized parameters in each fold. The table shows that even $J = 2$ yields reliable prediction of emotions. Notice that increasing the values of J significantly increases the number of elements in A^{s^*}, and, consequently, the number of features.

Table 5 reports the classification performance for the proposed blind lexical compensation method. The table lists the performance when we consider different combination of features (columns under "With Boundaries"). The first row corresponds to the baseline setting with no lexical compensation, which achieves 53.8% accuracy and 52.2% F-score. The evaluation shows that using A^{s^*} alone yields 58.5% accuracy. The improvement over the baseline lexical-independent method is statistically significant (p-value$< 1e-10$). Although the performance obtained by $\mathbf{d}^{s^*}$ is not competitive compared with other settings, it is significantly higher than a random classifier (25% chance), which demonstrates the discriminative power of this feature set. Fusing the feature sets results in performances above 60% accuracy, which shows a statistically significant improvement over the baseline performance (p-value$< 1e-10$). Metallinou et al. [51] used 27 facial markers in the lower and middle facial area of the IEMOCAP database for this four-emotion classification task. The study reported 54.2% average recall. Although

TABLE 6
Confusion matrices of the baseline classifier (first row in Table 5) and the classifier trained with all the features (last row in Table 5) – IEMOCAP.

	$\mathbf{l}^{bas}$ [baseline]				$\mathbf{l}^{bas}$, $\mathbf{d}^{s^*}$, A^{s^*}			
	Hap	Ang	Sad	Neu	Hap	Ang	Sad	Neu
Hap	594	142	58	33	619	93	60	55
Ang	161	309	42	62	96	359	31	88
Sad	109	83	309	125	99	23	345	159
Neu	103	137	151	193	105	89	140	250

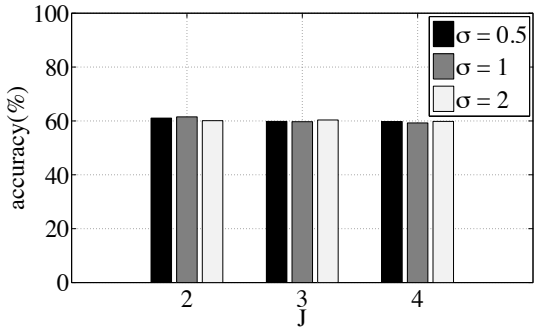


Fig. 7. The performance of the blind lexical compensation method across different values of J and σ.

our approach only uses 15 out of these 27 markers, the blind lexical compensation yields 58.8% average recall, which is a statistically significant improvement (p-value< 0.0003). Table 6 reports the confusion matrices of the baseline classifier, and the classifier trained with all the features (first and last row in Table 5, respectivily – "With Boundaries" condition). Table 6 shows performance improvements for all the emotional classes. The proposed lexical compensation is particularly effective in reducing the confusion between happiness and anger. After lexical compensation, the most common mistake is between sadness and neutral speech, reflecting the results in Figure 5.

5.1.2 Parameter Sensitivity Analysis

Although the optimized parameters reported in Table 4 vary across folds, they are not very different from each other ($J \in \{2, 3, 4\}$, $\sigma \in \{0.5, 1, 2\}$). In fact, the performance of the system is not very sensitive to the values for J (i.e., dimension of the subspace of the style matrix A^s) and σ (i.e., the standard deviation associated with the Gaussian distribution in Eq. 8). To demonstrate this finding, we retrain the system for different values of J and σ across folds. We only consider the classifier trained with all the features $\mathbf{l}^{bas}$, $\mathbf{d}^{s^*}$, and A^{s^*} (i.e., last row in Table 5). Figure 7 shows the performance for various combination of values assigned to J and σ. The figure shows consistent performance across different values of the parameters. The low value of J implies that the dimension of the subspace to characterize emotions can be small and still provides good discrimination between emotions.

TABLE 7
Emotion recognition results for supervised lexical compensation methods, by fusing the viseme-level labels - IEMOCAP (results from Mariooryad & Busso [12]).

Lexical compensation	Accuracy [%]	F-Score [%]
Feature level	56.26	54.16
Model level	**58.37**	**56.66**

5.1.3 Blind Lexical Compensation Without Phoneme Boundaries

Although the viseme labels are assumed to be unknown, the proposed blind lexical compensation approach uses the temporal phonetic boundaries (only the viseme segmentation, not the viseme labels). However, in practice this information is not available. We demonstrate that the proposed approach provides competitive performance even when the phonetic boundaries are not available. For this purpose, we randomly segment the testing utterances into short segments, and we implement the blind lexical compensation approach using the random viseme segmentation. The average duration of the visemes in the database is 104 ms. Hence, we assume an exponential distribution with mean duration equals to 104 ms. We obtain the viseme boundaries by sampling this distribution. Notice that viseme boundaries and their labels are used to train the bilinear model.

The last two columns of Table 5 report the results of this experiment (columns under "Random Boundaries"). Although we observe a decrease in performance compared to the case where ground truth viseme boundaries are given, the drop in performance is not statistically significant (p-value> 0.05). This result indicates that even without accurate viseme boundaries the proposed model can compensate for the lexical variability.

5.1.4 Supervised Lexical Compensation

In Mariooryad and Busso [12], we proposed two supervised lexical compensation methods (feature-level and model-level), which assume that viseme labels and their time alignment information are both available for training and testing the models (a limitation that this study overcomes). This section presents these methods and their performances as a reference. Both methods makes use of the viseme-level emotion recognizers, so we use the sum of the class likelihoods to derive a utterance level decision.

Feature-Level Supervised Lexical Compensation: The feature-level lexical compensation approach utilizes a viseme-level trajectory model to compensate for lexical variability. The lexical dependency in each viseme is reduced by applying viseme-dependent whitening transformations which unifies their distributions. Consider a given facial marker along a given direction. First, we build viseme-dependent trajectory models using all the instances of a given viseme in the training set. To build the models, we interpolate and resample the trajectories

TABLE 8
Arousal and valence classification results for the proposed blind lexical compensation method – SEMAINE. The table reports results for speaking and non-speaking segments for the "users" ($K = 2, 3$).

	Speaking segments								Non-speaking segments							
	K=2				K=3				K=2				K=3			
	Arousal		Valence		Arousal		Valence		Arousal		Valence		Arousal		Valence	
Feature Sets	A[%]	F[%]	A[%]	F[%]	A[%]	F[%]	A[%]	F[%]	A[%]	F[%]	A[%]	F[%]	A[%]	F[%]	A[%]	F[%]
$\mathbf{l}^{bas}$ [baseline]	53.81	52.18	66.38	45.38	31.1	24.79	39.13	27.42	44.59	37.96	68.13	47.61	31.13	23.91	28.61	23.86
$\mathbf{d}^{s^*}$	43.62	41.36	65.11	48.00	34.82	27.38	47.61	36.03	62.16	56.58	67.03	49.74	41.67	34.55	44.25	33.61
A^{s^*}	58.52	57.29	68.08	56.05	43.24	37.06	50.58	45.94	64.74	61.36	66.42	53.73	41.18	34.77	45.97	34.5
$\mathbf{l}^{bas}$, $\mathbf{d}^{s^*}$	54.58	52.92	64.16	47.96	28.65	22.41	38.18	28.59	46.07	39.36	66.67	50.75	37.13	30.54	39.00	34.53
$\mathbf{l}^{bas}$, A^{s^*}	60.36	59.10	65.32	54.25	44.62	37.77	47.72	40.42	63.51	59.60	66.06	53.95	42.77	38.00	43.77	35.47
$\mathbf{d}^{s^*}$, A^{s^*}	58.83	57.54	66.49	53.93	43.66	37.45	50.48	47.31	64.62	60.8	65.93	49.22	44.61	39.68	45.60	35.69
$\mathbf{l}^{bas}$, $\mathbf{d}^{s^*}$, A^{s^*}	60.25	58.94	65.22	53.84	44.83	38.5	48.78	38.46	64.25	60.42	66.18	49.73	45.47	40.45	45.60	36.86

to represent their temporal shape with a $N \times 1$ trajectory vector. As a result, the trajectory model of marker M (e.g., central chin marker – Fig. 3(c)), in a given direction (e.g., up-down movement), for viseme V is parametrized with its mean vector μ_{MV} and its covariance matrix Σ_{MV}. We compensate for the lexical information by using the whitening transformation. Let's consider the eigenvalue decomposition of $\Sigma_{MV} = U_{MV}\Lambda_{MV}U_{MV}^{-1}$. The whitening transformation is an affine linear transformation, where the output random vector is zero mean and has the identity as its covariance matrix. The transformation is given in Equation 14.

$$x^w = \Lambda_{MV}^{-\frac{1}{2}} U'_{MV}(x - \mu_{MV}) \tag{14}$$

Notice that a different transformation is applied to each viseme class V. Therefore, the lexical dependency on the trajectory of the facial features is reduced. The remaining signal conveys other factors, including emotions. Motivated by our previous works [55], the normalization parameters are estimated from the neutral portion of the database. This approach ensures that the distributions of the normalized trajectories in the neutral samples are similar, regardless of the underlying phonetic content (i.e., a zero mean random vector with identity covariance). Deviation from these statistics signal deviations in the facial features from neutral emotion.

Model-Level Supervised Lexical Compensation: The model-level lexical compensation method consists in training one emotion classifier for each of the viseme classes. Therefore, all the instances processed by a given viseme-dependent classifier will convey similar lexical information. Since we consider 13 visemes, the approach requires to train 13 different classifiers. The drawback of the approach is that the training data per classifier is reduced.

Table 7 reports the utterance level recognition rates for the feature level and models level supervised lexical compensation approaches. Compared to the baseline without lexical compensation ($\mathbf{l}^{bas}$ in Table 5), the feature and model level methods yield 2.5% (relatively 4.6%) and 4.6% (relatively 8.5%) accuracy improvements, respectively. These are statistically significant improvements over the baseline setting for both methods (p-value< 0.038).

Similar to the results of the blind lexical compensation method, these results also demonstrate the importance of compensating for the lexical variability to improve the classification performance. An important limitation of these methods is that the transcription and the phonetic alignment is required, which reduces its usability in real applications. The proposed blind lexical compensation approach overcomes this problem. Notice that the results obtained with the blind lexical compensation method (see Table 5) even outperforms the performances of the supervised lexical compensation methods (see Table 7), although the differences are not statistically significant. The blind estimation of A^{s^*} simultaneously takes into account all the segmented visemes to find the best subspace. This simultaneous estimation is not leveraged by the supervised methods. Hence, we achieve higher performance even without viseme labels.

5.2 Evaluations on the SEMAINE Database

The analysis and classification experiments consider facial features derived from motion capture data. These features are not available in real-world applications. Therefore, this section demonstrates the effectiveness of the proposed method using facial features automatically derived from video recordings. We replicate the experiments with the SEMAINE database using the discrete binary classes for arousal and valence defined with the k-means algorithm (Sec. 3.2). We use a nine-fold LOSO cross-validation evaluation. In each fold, four speakers from the training set are used to build the bilinear model and lexical-independent viseme-level classifier. The remaining four speakers in the training set are used to build the turn-level classifier. We set the parameters of the bilinear model to the most common combination observed in Table 4 ($J = 2$, $\sigma = 2$). We use the 13 AUs described in Section 3.2 as image-based features. The frame-level AUs are sampled at 25 Hz. However, the average duration of the visemes is approximately 100ms. Therefore, the seven high level statistics used for the IEMOCAP do not yield robust features. Instead, we

only estimate the mean, maximum and minimum of the AUs during each viseme, creating a 39-D feature vector (13 AUs $\times$ 3 statistics).

The columns under "speaking segments" in Table 8 presents the results on the SEMAINE database for arousal and valence, for segments when the "user" was speaking. In both emotional dimensions, the blind lexical compensation scheme significantly improves the performance. For valence with K=2, fusing the bilinear model features ($\mathbf{d}^{s^*}$, A^{s^*}) and the lexical-independent features ($\mathbf{l}^{bas}$) do not increase the performance. However, the F-score is higher than the lexical-independent baseline ($\mathbf{l}^{bas}$). The F-score improvements over the baseline classifier for all the conditions trained with features including A^{s^*} are statistically significant (p-value< 0.01).

5.2.1 Performance during Absence of Speech

The previous evaluations considered speaking turns. An interesting question is whether the performance of the proposed system is competitive when subjects are not speaking (i.e., non-speaking turns). We evaluate the performance of the blind lexical compensation method during the turns when the "users" were listening to their interlocutors. Instead of retraining the models under this condition, we evaluated the same bilinear models trained with speaking turns, assessing the generalization of the models. Since this evaluation considers silent segments, we do not have phonetic boundaries. We use the approach presented in Section 5.1.3, where the boundaries are randomly sampled using an exponential distribution.

The columns under "non-speaking segments" of Table 8 report the results of this evaluation. Similar to speaking segments, the proposed method significantly improves the performances over the baseline model. Although the participants are not speaking in these segments, they may still have facial movements unrelated to their emotional state. The bilinear model maps those movements to the closest viseme (i.e., content), finding the best emotional space for that segment. With the exception of the evaluation of valence with K=2, any condition containing A^{s^*} in the feature set improves the F-value over the baseline classifier. The improvements are statistically significant with p-values close to zero.

5.3 Comparison With other Methods

We compared our approach with the method proposed by Zeng et al. [33], [34], consisting in averaging the facial features over consecutive frames. This approach reduces the dependency of lexical information over facial features. Following their settings, we apply a moving averaging window to the facial features with a window size of 300 milliseconds. We implemented the baseline classifiers trained with this smoothing approach ($\mathbf{l}^{bas}$ - smoothing). The evaluation considers only speaking turns.

Tables 9 and 10 present the results for the IEMOCAP and SEMAINE databases, respectively. For comparison,

TABLE 9
Emotion recognition results with smoothing the facial feature to remove speaking effect [33], [34] - IEMOCAP.

	Accuracy [%]	F-Score [%]
$\mathbf{l}^{bas}$ [baseline]	53.8	52.2
$\mathbf{l}^{bas}$ - Smoothing	54.3	52.1
$\mathbf{l}^{bas}$, $\mathbf{d}^{s^*}$, A^{s^*}	60.3	58.9

TABLE 10
Emotion recognition results with smoothing the facial feature to remove speaking effect [33], [34] - SEMAINE.

	K=2				K=3			
	Arousal		Valence		Arousal		Valence	
	A [%]	F [%]	A [%]	F [%]	A [%]	F [%]	A [%]	F [%]
$\mathbf{l}^{bas}$ [baseline]	53.8	52.2	66.4	45.4	31.1	24.8	39.1	27.4
$\mathbf{l}^{bas}$ - Smoothing	54.0	46.9	64.2	46.4	34.7	27.2	32.9	24.9
$\mathbf{l}^{bas}$, $\mathbf{d}^{s^*}$, A^{s^*}	60.3	58.9	65.2	53.8	44.8	38.5	48.8	38.5

the tables also include the results presented in Tables 5 and 8 for the baseline classifiers ($\mathbf{l}^{bas}$), and the classifiers trained with all the features ($\mathbf{l}^{bas}$, $\mathbf{d}^{s^*}$, A^{s^*}). The evaluation shows that our approach outperforms this method. The p-value for the proportion hypothesis test is close to zero, showing significant improvements across all conditions. The smoothing approach does not provide significant improvements over the baseline classifier. In fact, in some cases the performance dropped.

6 Conclusions and Discussion

Given the challenges in processing facial expressions when the subject is speaking, this work introduced a novel approach for lexical compensation based on the asymmetric bilinear model. The proposed technique does not require phoneme labels, which is a significant improvement over the supervised lexical compensation methods presented in our previous work [12]. We evaluated the approach with emotion recognition experiments with controlled recordings, where motion capture data is available (IEMOCAP database), and with regular recordings, where AUs were automatically extracted from the videos (SEMAINE database). The results in both evaluations showed the benefits of the proposed blind lexical compensation technique to mitigate the effects of speech on facial expressions. Notice that in experiments on both databases, when A^{s^*} is included in the features, we achieve statistically significant improvements in F-value over the baseline classifier (the only exception is for the SEMAINE evaluation for valence, K=2, during silence segments - Table 8). The p-value for the proportion hypothesis test between the baseline framework and conditions trained with A^{s^*} is less than 0.01, and, in most of the cases, close to zero, showing significant improvements.

Our future research directions include reimplementing the approach with other facial features automatically derived from videos. We are considering to train the

asymmetric bilinear model with geometric features [56] and appearance-based features [57]. The experimental evaluation showed that the proposed approach achieves competitive performance even when random segmentation is used to estimate the viseme boundaries. The performance may increase if the phone boundaries are estimated using automatic algorithms [58].

Our analysis showed that the middle and upper facial areas are not affected by the spoken content. Hence, lexical-independent models are sufficient to characterize emotions. Given the promising results of the proposed lexical compensation scheme in the lower face region, we plan to develop an integrated system. The approach will incorporate lexical-independent models for the middle and the upper face regions, and lexical-dependent models for lower face region. We expect to obtain a facial expression recognition system that is robust against the variability introduced by the articulatory process.

ACKNOWLEDGMENTS

The authors would like to thank the Machine Perception Lab (MPLab) at UCSD for providing the CERT toolkit.

REFERENCES

[1] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 15, no. 8, pp. 2331–2347, November 2007.

[2] S. Mariooryad and C. Busso, "Factorizing speaker, lexical and emotional variabilities observed in facial expressions," in *Proc. IEEE Int. Conf. Image Processing*, Orlando, FL, USA, September-October 2012, pp. 2605–2608.

[3] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Expression-invariant 3D face recognition," in *Audio- and Video-Based Biometric Person Authentication*. Guildford, UK: Springer Berlin Heidelberg, June 2003, vol. 2688, pp. 62–70.

[4] B. Park, K. Lee, and S. Lee, "Face recognition using face-ARG matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 27, pp. 1982–1988, December 2005.

[5] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3D face recognition with a morphable model," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recog.*, Amsterdam, The Netherlands, September 2008.

[6] I. Mpiperis, S. Malassiotis, and M. Strintzis, "Bilinear models for 3-D face and facial expression recognition," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 3, pp. 498–511, September 2008.

[7] O. Rudovic, I. Patras, and M. Pantic, "Coupled Gaussian process regression for pose-invariant facial expression recognition," in *ECCV*. Heraklion, Crete, Greece: Springer Berlin Heidelberg, September 2010, vol. 6312, pp. 350–363.

[8] E. Vatikiotis-Bateson, K. Munhall, Y. Kasahara, F. Garcia, and H. Yehia, "Characterizing audiovisual information during speech," in *Proc. 4th Int. Conf. Spoken Lang. Process. (ICSLP 96)*, vol. 3, Philadelphia, PA, USA, October 1996, pp. 1485–1488.

[9] 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.

[10] Y. Kim and E. Mower Provost, "Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition," in *ACM Int. Conf. Multimedia*, Orlando, FL, USA, November 2014, pp. 27–36.

[11] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *Proc. 7th Int. Seminar Speech Production (ISSP06)*, Ubatuba-SP, Brazil, December 2006, pp. 549–556.

[12] S. Mariooryad and C. Busso, "Feature and model level compensation of lexical content for facial emotion recognition," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recog.*, Shanghai, China, April 2013.

[13] J. Tenenbaum and W. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, June 2000.

[14] J. B. Tenenbaum and W. T. Freeman, "Separating style and content," Cambridge, MA, Technical Report TR96-36, December 1996.

[15] P. Ekman, T. S. Huang, T. Sejnowski, and J. C. Hager, "Final report to NSF of the planning workshop on facial expression understanding," National Science Foundation, University of California, San Francisco, CA, USA, Technical Report, July-August 1992.

[16] M. Kamachi, V. Bruce, S. Mukaida, J. Gyoba, S. Yoshikawa, and S. Akamatsu, "Dynamic properties influence the perception of facial expressions," *Perception*, vol. 30, no. 7, pp. 875–887, July 2001.

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

[18] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," *Philos. Trans. R. Soc. B*, vol. 264, pp. 3505–3513, November 2009.

[19] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3D dynamic faces using robust spatio-temporal shape features," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recog.*, Santa Barbara, CA, USA, March 2011, pp. 414–421.

[20] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "A dynamic approach to the recognition of 3D facial expressions and their temporal models," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recog.*, Santa Barbara, CA, USA, March 2011, pp. 406–413.

[21] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *J. Image Vision Comput.*, vol. 30, no. 10, pp. 683–697, October 2012.

[22] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011- the first international audio/visual emotion challenge," in *ACII*. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, vol. 6975/2011, pp. 415–424.

[23] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Proc. Interspeech*, Portland, OR, USA, September 2012, pp. 1179–1182.

[24] R. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[25] M. Shah, D. Cooper, H. Cao, R. Gur, A. Nenkova, and R. Verma, "Action unit models of facial expression of emotion in the presence of speech," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interaction*, Geneva, Switzerland, September 2013, pp. 49–54.

[26] D. Datcu and L. Rothkrantz, "Semantic audiovisual data fusion for automatic emotion recognition," in *Emotion Recognition: A Pattern Analysis Approach*. Hoboken, NJ, USA: John Wiley & Sons, January 2015, pp. 411–436.

[27] L. Chen and T. Huang, "Emotional expressions in audiovisual human computer interaction," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME 2000)*, vol. 1, New York City, NY, USA, July-August 2000, pp. 423–426.

[28] Z. Yong, C. Yabi, and Z. Yongzhao, "Expression recognition method of image sequence in audio-video," in *Proc. 3rd Int. Symp. Intelligent Inf. Technol. Applicat.)*, vol. 1, Nanchang, China, November 2009, pp. 513–516.

[29] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *J. Image Vision Comput.*, vol. 26, no. 7, pp. 1052–1067, July 2008.

[30] J. Lin, C. Wu, and W. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 142–156, October 2012.

[31] J. N. Bassili, "Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face," *J. Personality and Social Psychology*, vol. 37, no. 11, pp. 2049–2058, November 1979.

[32] H. Hoffmann, H. C. Traue, K. Limbrecht-Ecklundt, S. Walter, and H. Kessler, "Static and dynamic presentation of emotions in different facial areas: Fear and surprise show influences of temporal and spatial properties," *Psychology*, vol. 4, no. 8, pp. 663–668, August 2013.

[33] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T. Huang, D. Roth, and S. Levinson, "Bimodal HCI-related affect recognition," in *Proc. Sixth ACM Intl Conf. Multimodal Interfaces*. State College, PA, USA: ACM Press, October 2004, pp. 137–143.
[34] Z. Zeng, J. Tu, M. Liu, T. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 424–428, February 2007.
[35] C. Wu, W. Wei, J. Lin, and W. Lee, "Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1732–1744, July 2013.
[36] 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," *J. Language Resource Eval.*, vol. 42, no. 4, pp. 335–359, December 2008.
[37] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, January-March 2012.
[38] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Dallas, TX, USA, March 2010, pp. 2474–2477.
[39] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," in *Proc. 2nd Int. Workshop Emotion: Corpora for Research on Emotion and Affect, Int. Conf. Lan- guage Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 1–8.
[40] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *Proc. ISCA Workshop Speech and Emotion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 19–24.
[41] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, April-June 2015, special Issue Best of ACII.
[42] ——, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 183–196, April-June 2013.
[43] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.
[44] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multimedia*, vol. 1, no. 6, pp. 22–35, September 2006.
[45] P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Proc. 10th Australian Int. Conf. Speech Science and Technology*, Sydney, NSW, Australia, December 2004, pp. 265–270.
[46] A. Solomonoff, C. Quillen, and W. Campbell, "Channel compensation for SVM speaker recognition," in *Odyssey04*, Toledo, Spain, May-June 2004, pp. 57–62.
[47] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annu. Eugenics*, vol. 7, no. 2, pp. 179–188, September 1936.
[48] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th Intl Conf. Computer Vision*, vol. Rio de Janeiro, Brazil, October, October 2007, pp. 1–8.
[49] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Montréal, QC, Canada, Technical Report CRIM-06/08-14, 2006.
[50] S. Yan, H. Wang, J. Tu, X. Tang, and T. Huang, "Mode-kn factor analysis for image ensembles," *IEEE Trans. Image Process.*, vol. 18, no. 3, pp. 670–676, March 2009.
[51] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Dallas, TX, USA, March 2010, pp. 2462–2465.
[52] E. Mower and S. Narayanan, "A hierarchical static-dynamic framework for emotion classification," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Prague, Czech Republic, May 2011, pp. 2372–2375.
[53] E. Provost and S. Narayanan, "Simplifying emotion classification through emotion distillation," in *APSIPA Annual Summit and Conference*, Hollywood, CA, USA, December 2012, pp. 1–4.
[54] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, June 2009.
[55] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Proc. Interspeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.
[56] F. Bourel, C. Chibelushi, and A. Low, "Robust facial expression recognition using a state-based model of spatially-localised facial dynamics," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recog.*, Washington, DC, USA, May 2002, pp. 106–111.
[57] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 974–989, October 1999.
[58] O. Kalinli, "Automatic phoneme segmentation using auditory attention features," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Kyoto, Japan, March 2012, pp. 2270–2273.

Soroosh Mariooryad (S'2012) received his B.S degree (2007) with high honors in computer engineering from Ferdowsi University of Mashhad, and his M.S degree (2010) in computer engineering (artificial intelligence) from Sharif University of Technology (SUT), Tehran, Iran. He received his Ph.D. degree (2014) in Electrical Engineering at the University of Texas at Dallas (UTD), Richardson, Texas, USA. He is currently at Google. From 2008 to 2010, he was a member of the Speech Processing Lab (SPL) at SUT. In 2010, he joined as a research assistant the Multimodal Signal Processing (MSP) laboratory at UTD. In summer 2013, he interned at Microsoft Research working on analyzing speaking style characterstics. His research interests are in speech and video signal processing, probabilistic graphical models and multimodal interfaces. His current research includes modeling and analyzing human non-verbal behaviors, with applications to speech-driven facial animations and emotion recognition. He has also worked on statistical speech enhancement and fingerprint recognition.

Carlos Busso (S'02-M'09-SM'13) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests include digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human interaction, in-vehicle active safety system, and machine learning methods for multimodal processing. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.