STYLE EXTRACTOR FOR FACIAL EXPRESSION RECOGNITION IN THE PRESENCE OF SPEECH

Ali N. Salman and Carlos Busso

Multimodal Signal Processing (MSP) laboratory, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
ans180000@utdallas.edu, busso@utdallas.edu

ABSTRACT

The performance of *facial expression recognition* (FER) systems has improved with recent advances in machine learning. While studies have reported impressive accuracies in detecting emotion from posed expressions in static images, there are still important challenges in developing FER systems for videos, especially in the presence of speech. Speech articulation modulates the orofacial area, changing the facial appearance. These facial movements induced by speech introduce noise, reducing the performance of an FER system. Solving this problem is important if we aim to study more naturalistic environment or applications in the wild. We propose a novel approach to compensate for lexical information that does not require phonetic information during inference. The approach relies on a style extractor model, which creates emotional-to-neutral transformations. The transformed facial representations are spatially contrasted with the original faces, highlighting the emotional information conveyed in the video. The results demonstrate that adding the proposed style extractor model to a dynamic FER system improves the performance by 7% (absolute) compared to a similar model with no style extractor. This novel feature representation also improves the generalization of the model.

Index Terms— Affective computing, facial expression recognition, style and content, speech articulation, factor analysis

1. INTRODUCTION

Perceiving the emotion in others is a crucial skill in human interaction. We effortlessly sense cues to estimate the expressed emotion of the interlocutor by relying on various modalities, especially information conveyed by face and speech. This skill is also useful in *human computer interaction* (HCI) with applications in improving areas such as automated tutoring systems [1, 2], human robot interaction [3], and detection of driver distractions [4, 5]. While we can unconsciously infer the emotions of others, this task is a real challenge for computers. This study focuses on *facial expression recognition* (FER) from videos while the target subjects are speaking.

Earlier studies on emotion classification from visual data have mostly focused on using a single image (snapshot) for classification. FER systems using videos is an important need due to the popularity of videos that are uploaded in social media everyday, and new applications that this technology can enable. However, there are several challenges to create a robust FER system using videos. The straightforward approach is to independently process each frame using an image-based FER. However, our previous study highlighted the intrinsic limitations of processing a sequence of images frame-by-frame without considering their temporal information [6]. While some studies have argued that a single image taken at the apex of an expression is optimal to detect emotions [7], finding meaningful apex frames is not easy. This process is even more complex in the presence of speech. When a subject speaks, articulatory movements affect the appearance of the face, especially in the orofacial area [8, 9]. The lexical information introduces nuances, reducing the reliability and performance of an emotion recognition system [10]. Solutions to compensate for the lexical variability include ignoring facial features in the orofacial area [11, 12], processing only silent frames [13, 14], and smoothing facial features to remove articulation [15]. These approaches are suboptimal since they ignore important information. A better solution is to develop lexical compensation methods that can reduce the effects of speech articulation in FER systems [16–18].

This paper proposes a novel approach to reduce the impact of speech articulation in FER problems. The key feature of the approach is to create transformations from emotional to neutral faces (e.g., happiness to neutral state). The style extractor transformation is implemented using a *deep learning* (DL) model that are trained with paired data conveying the same lexical content, but with different emotions. The transformed faces are contrasted with the original faces creating a discriminative feature representation that emphasizes the spatial deviations between emotional and neutral faces for the given sentence. During inference, the approach does not need phonetic information, operating as a blind lexical compensation FER approach. These discriminative lexical compensation features are combined with embeddings extracted from frames using the VGG16 network. For a four class problem (i.e., happiness, anger, sadness, and neutral state), the addition of our proposed style extractor model increases the F1-score by 7% (absolute) with respect to a model relying only on image embedding information. The results validate the effectiveness of our novel method to capture lexical-independent features extracted from image sequences without the need of costly transcriptions during inference.

2. BACKGROUND

2.1. Related Work

Studies have reported important advances in FER [19,20], especially with new architectures in DL. FER tasks usually consist of either *action unit* (AU) recognition or categorical emotion classification. This study focuses on categorical emotion classification (e.g., happiness, anger, sadness). Most of the studies have focused on image based classification, achieving good performance when the expression in the image is clear. However, building an FER system for videos is still a challenging problem. Studies have found that important temporal characteristics are lost when dynamic information is ignored [21, 22]. For example, temporal information provides valuable cues for subtle emotions [22, 23]. Salman and Busso [6] compared human performance while evaluating static representation of

This work was supported by NEC Foundation, and NSF under Grant IIS-1718944.

emotions in the presence of speech. The study showed that the emotion perception of isolated frames in a video is clearly different from the emotional perception after watching the corresponding videos. The differences also vary across emotions, where the perception of sadness and anger seems to depend more on the dynamic information. The implications of these studies indicate that analyzing images frame-by-frame in a video is not an ideal approach. Even if we have an image-based FER system with human-level performance, ignoring temporal information limits the ability to properly unveil the perceived emotion in a video.

Hoffmann *et al.* [24] studied the effects of different facial regions in the perception of emotions. They partitioned the face into upper and lower regions. They found that in dynamic visual sequences some emotions were better perceived in the lower region (i.e., happiness, anger, sadness, and disgust), so ignoring the orofacial area is not an ideal approach. Busso and Narayanan [25] partitioned facial motion capture markers into lower, middle, and upper regions on the recordings of an actress. The study showed that during speech the lower area had on average more activity in angry sentences compared to sad sentences. Although the lower area is greatly affected by speech articulation, this area is still informative for emotion recognition and should be considered in FER systems. Since the orofacial region conveys important cues in the expression of emotions, it is essential to develop FER algorithms that can distinguish between emotional facial information and lexical facial information associated with speech articulation. A straightforward approach to reduce the lexical variability is to train phone or viseme dependent FER classifiers [18, 26, 27]. Kim and Mower Provost [28] separated motion captured data of the face into upper and lower regions. Since the lower facial region is more deponent on the speech articulation, they used phone-dependent classifiers. Kim and Mower Provost [16] also showed that using phone-based segmentation and classification approach was better than unsupervised and window-slide based segmentation. The use of phone-dependent models requires transcriptions with aligned temporal information, which can be obtained using an *automatic speech recognition* (ASR) system. Other methods have attempted to reduce the effects of speech articulation on visual features without depending on phonetic alignment, creating blind-lexical compensation formulations [17, 28]. Marioooryad and Busso [17] used a bi-linear model to extract emotion dependent features. Our proposed framework is different from these methods, since our novel deep learning formulation directly estimates deviations from expected neutral facial movements, which are then used to compensate for lexical facial information.

2.2. Databases

The study uses the AffectNet [29], MSP-IMPROV [30], and CREMA-D [31] corpora, which are described in this section.

AffectNet: We use the AffectNet corpus [29] to train the feature extractor model, which consists of an image-based FER system (Sec. 3.1). AffectNet is a facial expression database consisting of over 1 million images collected from the internet. This database includes images instead of videos. Around 440 thousand images are manually annotated with seven discrete emotional labels. The annotations also included the emotional attributes valence and arousal, but these descriptors are not used in this study. The average resolution of the images is 425×425. We used a subset of this database for our experiments, where we only consider four emotions (happiness, anger, sadness, and neutral state). We randomly select 24,882 images for each emotion from the training set in the AffectNet corpus to achieve a balanced dataset. Because the test set is not publicly available, we further split this set into train (80%) and validation (20%) sets. We use the validation set of the AffectNet corpus as our test set. This set is balanced with 500 images per emotion.

MSP-IMPROV: We use the MSP-IMPROV corpus [30] to train our style extractor model (Sec. 3.2). The MSP-IMPROV corpus is a multimodal database collected in dyadic interactions from 12 subjects (six females and six males). The videos were collected in a controlled environment at 29.97fps with a resolution of $1{,}440 \times 1{,}080$ pixels. The corpus was created to study multimodal emotion perception [32]. The objective in the corpus was to have multiple target sentences conveying four emotions as naturally as possible (happiness, anger, sadness and neutral). A hypothetical scenario was created per sentence and per emotion. The scenario led one of the actors to speak the target sentence in the target emotion. The database consists of the target sentences, the rest of the speech segments in the improvisation and natural conversations between the actors during the break. We only use the target sentences, which were annotated using *Amazon Mechanical Turk* (AMT) by multiple workers. We separately annotated the video-only, audio-only and audiovisual conditions using categorical emotions [32]. This study considers the annotations of the video-only condition, using only the videos where a consensus agreement is reached for the emotions happiness, anger, sadness, and neutral state (majority vote rule).

CREMA-D: We use the CREMA-D corpus to train and test the fusion model in our proposed method after pre-training the style extractor and feature extractor models (Sec. 3.3). The CREMA-D corpus [31] is another multimodal emotional database with sentences with predetermined lexical content in different emotions. The corpus is an audiovisual dataset purposely recorded for emotion recognition and perception. The recordings were collected in a controlled environment with a resolution of 960×720 pixels. Professional actors recorded eleven target sentences conveying happiness, anger, sadness, fear, disgust, and neutral state. One additional sentence was acted with three intensity levels for each emotion, resulting in 18 clips per actor. For each clip, the audio, visual, and audiovisual representations were separately annotated using a crowdsourcing evaluation. The corpus consists of 7,442 annotated clips from 91 professional actors. Our study uses the consensus labels obtained using the majority vote rule of the video-only annotations. We only consider the emotions happiness, anger, sadness and neutral state, resulting in a set of 5,093 videos. We use data from 81 actors for our train set, data from four actors for the validation set, and data from six actors for the test set. The validation and test sets are gender balanced.

3. PROPOSED APPROACH

Our proposed FER framework aims to reliably predict the perceived emotion in the videos while a subject is speaking. We achieve this goal by creating emotional to neutral facial transformations that are temporally and spatially contrasted with the original emotional faces. The differences convey expressive information, compensating for the lexical facial information. This approach is a blind-lexical compensation formulation that does not require phonetic information during inference. Figure 1 shows the proposed model which consists of three parts. The first block is the *feature extractor model*, which produces facial features from the original images using multiple layers of *convolutional neural network* (CNN). The second block is the *style extractor model*, which is the core contribution of this study. This block compensates for the lexical variability, overcoming the noise introduced by speech articulation. The third block is the *fusion model*, which concatenates the feature representations from the first two models to predict emotions.

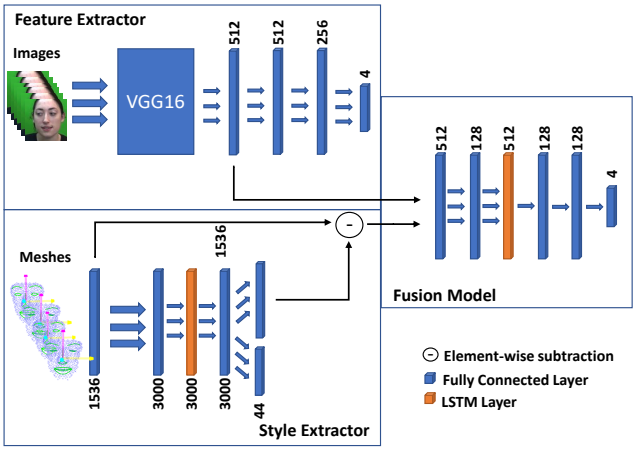

Fig. 1. Diagram of the proposed FER system for videos. The key contribution in this study is the style extractor, which aims to compensate for lexical information.

3.1. Feature Extraction Model

The purpose of this model is to extract a discriminative facial feature representation directly from images (Fig. 1). This vector will be combined with the feature representation produced by the style extractor model (Sec. 3.2). The feature extraction model is static, processing each frame without considering temporal information (the style extraction model and the fusion model will incorporate temporal information). We use the VGG16 architecture [33] for our model. After the VGG16 network, we add three dense layers of 512, 512, and 256 nodes, respectively. The output layer for this block is a softmax layer with four neurons to predict the emotion in the images (happiness, anger, sadness, neutral state). The loss function for this model is the categorical cross-entropy loss. The feature representation that is passed to the fusion model is the first dense layer after the VGG16 max pooling layer (512 nodes, Fig. 1).

3.2. Style Extractor Model

The style extractor model creates a facial transformation from emotional (i.e., happiness, anger, sadness) to neutral state for each frame. Then, the transformed images are used as neutral reference to temporally and spatially contrast the original emotional images. The resulting vector is expected to convey emotional information after compensating for lexical information.

Instead of operating with pixels as our input, we decide to implement the facial transformations relying on normalized 3D meshes with 512 points. This approach has two key advantages. First, we can significantly reduce the dimension of the feature vector. We implement the approach with 512 facial points, which reduces the dimensionality of the input to only 1,536 (i.e., 3×512). Second, the 3D meshes reduce the speaker dependent features and the noise due to background pixels, which is important as the intended facial transformation needs to be subject independent. We use the Zface toolkit [34] to extract the 3D facial mesh and train the style extractor model.

We train the style extractor model with paired data consisting of two aligned videos conveying the same lexical content, but with different emotions. Since we are interested in an emotional-to-neutral facial transformation, one of the video is emotional (happiness, anger or sadness), and the other is neutral, where the goal is to learn the mapping between both 3D meshes. The input of the model is the 3D mesh of the emotional face. We pass the input (3D meshes) to a shared dense layer of 3,000 neurons followed by a *long short-term memory* (LSTM) layer with 3,000 neurons. The LSTM layer is added to model temporal information. The LSTM output is then passed to another shared dense layer of 3,000 neurons. The output of the model is (1) the transformed neutral 3D mesh, which is a dense layer with 1,536 neurons, and (2) a softmax layer of 44 neurons, representing the one-hot encoding of the phone assigned to each mesh. The softmax layer facilitates learning phone-dependent features without requiring phonetic alignment during testing. The loss function is a linear combination of the mean squared error of the meshes, and categorical cross-entropy of the phones. Both losses are equally weighed. We use the difference between the emotional mesh (input) and predicted neutral mesh (output) as the style feature vector, which is then passed to the fusion model.

A key requirement to build this model is the aligned paired data of the meshes. We need to create image pairings where the speech articulatory information (i.e., phonetic content) is identical in both 3D meshes and the emotions (i.e., style) are different. We use the MSP-IMPROV corpus to create the pairings of images, since the target sentences provide videos where the lexical content is identical, but the perceived emotion is different. These sequences have to be aligned. We use the Montreal forced alignment toolkit [35], which takes the audio and the transcription, producing the timing information for each phone. The timing information is used to align the emotional and neutral frames, and to assign a phone to each frame in the video. The phone sequence for clips with the same lexical content may differ due to differences in pronunciation. We resolved these problems by manually correcting the alignment. The pairs are created by considering videos from the same subject. In total, we have 15,973 (happiness), 10,595 (sadness), and 10,720 (anger) paired images to train the models.

Table 1. Performance of the static FER system in the feature extractor model. The approach is implemented with the VGG16 network using the AffectNet corpus.

Emotion	Precision [%]	Recall [%]	F1-score [%]
Happiness	89.8	91.0	90.5
Anger	76.7	71.2	73.9
Sadness	75.8	71.6	73.7
Neutral	63.7	70.1	67.0
Average	76.5	76.2	76.3

3.3. Fusion Model

The fusion model concatenates the feature representations provided by the feature extractor and style extractor models along the time axis. The goal of the fusion model is to predict an emotion for the entire sequence of frames. Each frame is then passed into two shared dense layers of size 512 and 128, respectively. Then, we use a LSTM layer with 512 neurons to extract temporal information. The LSTM layer returns the last output of the sequence. This single feature vector is then passed to two dense layers with 512 and 128 neurons, respectively. The output layer is a softmax with four neurons, representing the emotional classes. The loss function is the categorical cross-entropy loss.

4. EXPERIMENTAL EVALUATION

4.1. Implementation

We use the train and validation set of the AffectNet database to build the feature extractor model. We initialize the weights of the VGG16 network with the VGG-Face weights provided by Parkhi *et al.* [36]. We use dropout with the rate equal to p=0.25. The training of the

Table 2. Evaluation of the emotional-to-neutral transformation in the style extractor model. The table shows the distribution of emotions recognized by an emotion classifier evaluated with original and transformed meshes. The classifier is trained with the MSP-IMPROV corpus and tested with the CREMA-D corpus.

Mesh	Happiness	Anger	Sadness	Neutral
Original	15,886	191,309	332,825	25,060
Transformed	3,332	122,075	260,350	179,323

feature extraction model has two parts. In the first part, we train the dense layers starting with random initialization using the pre-trained weights for the VGG16 network. Notice that the VGG-Face weights were optimized for face recognition problems, so they are not tuned for FER. Therefore, the second part of the training aims to tune the VGG16 weights toward our FER task. We jointly train the last CNN block of VGG16 with the dense layers. Table 1 shows the F-1 score of the model on our AffectNet testing set (Sec. 2.2). We achieve an average F1-score of 76.3% across emotions. This result corresponds to image classification on the AffectNet set. Section 4.2 presents FER results achieved by the full architecture on videos from the CREMA-D database.

We train the style extractor model with the MSP-IMPROV corpus, using the aligned paired meshes (Sec. 3.2). We consider sequences of 60 frames (i.e., 2 seconds) with a step size of 10 frames. We use dropout with a rate r=0.5 between the layers.

We freeze the weights of the feature extractor and the style extractor models after training these networks on the AffectNet and MSP-IMPROV, respectively. Then, we train the fusion model using the CREMA-D corpus. Notice that all the images and paired sequences to train the previous models come from the other corpora, avoiding having the same sequences in the test set. Similar to the MSP-IMPROV, we consider sequences of 60 frames.

The proposed architecture is implemented with Keras and TensorFlow. All the models are implemented with ADAM optimizer, with a learning rate equal to r=0.001. We use batch normalization, ReLU activation for dense layers, and tanh for the LSTM layer.

4.2. Results

The first part of the evaluation considers the effectiveness of the style extractor model with a discriminative analysis. The key idea of this evaluation is that emotional 3D meshes should look more "neutral" after the transformation. We train a simple network to classify emotion using the 3D meshes as input. The network is implemented with three fully connected layers with 512, 256 and 128 neurons, respectively. We train the models with the original 3D meshes obtained from the train set of the MSP-IMPROV corpus. As a reference, the model achieves an F1-score of 60.1% on our MSP-IMPROV test set. To evaluate the style extractor model, we classify the emotions of the original meshes from the CREMA-D corpus. Only 4.4% of meshes are classified as neutral. We also classify the emotions of the transformed meshes created by the style extractor model. For the transformed meshes, we found that 31.7% of these meshes are predicted as neutral. This shift in the distribution indicates that our emotional-to-neutral transformation is effective. Table 2 shows the number of frames assigned to each class for the original and transformed meshes.

The second part of the evaluation assesses the effectiveness of our proposed approach. We evaluate our architecture with and without the style extractor. Notice that both approaches include temporal information, since the fusion model is implemented with LSTM units. Table 3 shows the F1-score of this model with and without the style extractor. The results clearly demonstrate the effectiveness of using the style extractor model, which leads to higher F1-score on all the emotions. On average, adding the style extractor improves the F1-score achieved by the FER model from 67% to 74%, which corresponds to a 7% (absolute) gain. The improvement is directly attributed to the additional features provided to the fusion model by the style extractor model. It is also worth noting that while the gap in performance on the training set is much closer (within 1-2%), the model with the style features achieves a higher accuracy on the validation and testing sets. Therefore, the features from the style extractor model also contribute to improve the generalization of the FER models.

Table 3. Performance of the proposed FER system for videos on the test set of the CREMA-D corpus. Model A includes the style extractor. Model B does not include the style extractor.

Emotion	Precision		Recall		F1-score	
Model	A [%]	B [%]	A [%]	B [%]	A [%]	B [%]
Happiness	87.8	81.1	83.0	83.5	85.3	82.3
Anger	89.2	51.0	50.9	65.0	64.8	57.1
Sadness	78.6	83.0	60.5	52.3	68.4	64.1
Neutral	68.8	65.0	89.9	65.0	78.0	65.0
Average	81.1	70.0	71.0	66.4	74.1	67.1

5. CONCLUSIONS

This study proposed a novel deep learning method for FER in the presence of speech that does not require motion captured data or phonetic information during inference. The key contribution of this study is the style extractor model which relies on sequences of phonetically paired image data to estimate an emotional-to-neutral facial transformation. The style extractor uses LSTM to dynamically normalize the emotional facial features retaining the lexical based articulations. The transformed facial meshes are used as a reference to spatially contrast the emotional content observed after compensating for the lexical information. The proposed lexical-independent feature representation is concatenated with facial features directly extracted from a static FER system. We found that the features from the style extractor model improve not only the FER performance, but also the generalization of the model.

For future research, we will explore different feature extraction and fusion models to improve performance. Additionally, we would like to improve and automate the alignment process which is currently a laborious process. While the paired data was able to improve accuracy and generalization for FER tasks, we expect that our approach can also be useful in other tasks. The approach can be useful in image-to-image transformations using *generative adversarial networks* (GANs), where the goal is to transform facial appearance from one emotion to another while preserving the lexical content. The paired data can also be used to analyze how facial features are affected in the presence of speech to help produce realistic facial animation.

6. REFERENCES

[1] J. Whitehill, M. Bartlett, and J. Movellan, "Automatic facial expression recognition for intelligent tutoring systems," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2008)*, Anchorage, AK, USA, June 2008, pp. 1–6.

[2] A. Sarrafzadeh, S. Alexander, F. Dadgostar, C. Fan, and A. Bigdeli, "See me, teach me: Facial expression and gesture recognition for intelligent tutoring systems," in *Innovations in*

Information Technology (IIT 2006), Dubai, United Arab Emirates, November 2006, pp. 1–5.

[3] D. Kulic and E. A. Croft, "Affective state estimation for human-robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, October 2007.

[4] N. Li and C. Busso, "Analysis of facial features of drivers under cognitive and visual distractions," in *IEEE International Conference on Multimedia and Expo (ICME 2013)*, San Jose, CA, USA, July 2013, pp. 1–6.

[5] ——, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 51–65, February 2015.

[6] A. Salman and C. Busso, "Dynamic versus static facial expressions in the presence of speech," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, Buenos Aires, Argentina, May 2020.

[7] J. Gold, J. Barker, S. Barr, J. Bittner, W. Bromfield, N. Chu, R. Goode, D. Lee, M. Simmons, and A. Srinath, "The efficiency of dynamic and static facial expression recognition," *Journal of Vision*, vol. 13, no. 5, pp. 1–12, April 2013.

[8] C. Busso and S. Narayanan, "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.

[9] 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.

[10] 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 *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 49–54.

[11] 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.

[12] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, September 2003.

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

[14] Z. Yong, C. Yabi, and Z. Yongzhao, "Expression recognition method of image sequence in audio-video," in *International Symposium on Intelligent Information Technology Application (IITA 2009)*, vol. 1, Nanchang, China, November 2009, pp. 513–516.

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

[16] Y. Kim and E. Mower Provost, "Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition," in *ACM International Conference on Multimedia (MM 2014)*, Orlando, FL, USA, November 2014, pp. 27–36.

[17] S. Mariooryad and C. Busso, "Facial expression recognition in the presence of speech using blind lexical compensation," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 346–359, October-December 2016.

[18] ——, "Feature and model level compensation of lexical content for facial emotion recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013)*, Shanghai, China, April 2013, pp. 1–6.

[19] S. Li and W. Deng, "Deep facial expression recognition: A survey," *ArXiv e-prints (arXiv:1804.08348)*, pp. 1–25, April 2018.

[20] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 325–347, July-September 2019.

[21] D. W. Cunningham and C. Wallraven, "Dynamic information for the recognition of conversational expressions," *Journal of Vision*, vol. 9, no. 13, pp. 1–17, December 2009.

[22] 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, May 2005.

[23] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," *Philosophical Transactions of Royal Society B*, vol. 264, pp. 3505–3513, November 2009.

[24] 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.

[25] 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.

[26] 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.

[27] Y. Kim and E. Mower Provost, "Emotion recognition during speech using dynamics of multiple regions of the face," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 15, pp. 25:1–25:23, October 2015.

[28] ——, "ISLA: Temporal segmentation and labeling for audio-visual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 196–208, April-June 2019.

[29] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, January-March 2019.

[30] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[31] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

[32] E. Mower Provost, Y. Shangguan, and C. Busso, "UMEME: University of Michigan emotional McGurk effect data set," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 395–409, October-December 2015.

[33] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4845–4849.

[34] L. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D videos in real-time," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015)*, Ljubljana, Slovenia, May 2015, pp. 1–8.

[35] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 498–502.

[36] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC 2015)*, Swansea,UK, September 2015, pp. 1–12.