Facial Emotion Recognition in Presence of Speech using a Default ARTMAP Classifier

Sheir Afgen Zaheer[1,2] and Jong-Hwan Kim[1]
[1]*School of Electrical Engineering, KAIST, Daejeon, Republic of Korea*
[2]*Innovative Play Lab, Goyang, Republic of Korea*

Keywords: Emotion Recognition, Fuzzy Adaptive Resonance Theory, Default ARTMAP.

Abstract: This paper proposes a scheme for facial emotion recognition in the presence of speech, i.e. the interacting subjects are also speaking. We propose the usage of default ARTMAP, a variant of fuzzy ARTMAP, as a classifier for facial emotions using feature vectors derived from facial animation parameters (FAP). The proposed scheme is tested on Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. The results show the effectiveness of the approach as a standalone facial emotion classifier as well as its relatively superior performance on IEMOCAP in comparison to the existing similar approaches.

1 INTRODUCTION

To realize emotional intelligence in robots and artificial intelligence, ability to process emotional information and recognize emotions is essential. People communicate their emotions through various modes of communication. Facial expressions are the most dominant indicators of emotions among those communication cues. Therefore analyzing facial information for emotion recognition has attracted a lot of interest as research issue various fields, such as affective computing, social robotics and human robot interaction (Liu et al., 2013; Hirota and Dong, 2008; Rozgi et al., 2012).

In recent years, machine learning techniques for facial emotion recognition have been very popular (Liu et al., 2014; Li et al., 2015b). Among those Convolutional Neural Networks (CNN) have been the most successful and popular on the benchmark problems (Li et al., 2015a). These approaches use the images or sections of the images directly as training inputs. Though such approaches have been very successful on popular facial emotion databases, such as MMI and CKP facial expression database, they have practical limitations with audiovisual data consisting of multi-modal interactions. They work really well for still image data or video data with facial expressions only. However, this changes when the incoming data is audiovisual and the user is speaking. The variations in a speaking face are a compound effect of both the facial expression (emotion) and the facial movement to utter the words (lexicon).

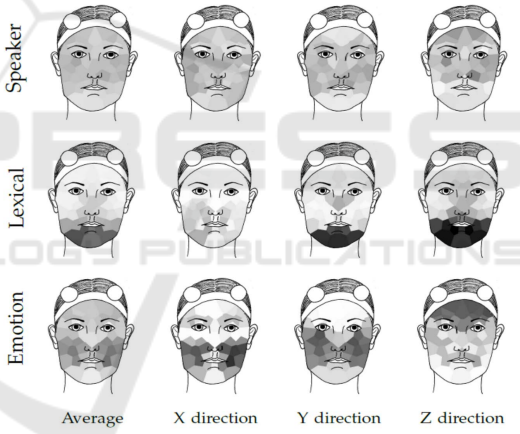

Figure 1: Dependency of various factors on different areas of the face. Darker color represents higher dependency.

To extract right features for facial emotion from a speaking face, we first need to understand how the emotions and lexicon affect different regions of the face. (Mariooryad and Busso, 2016) studied this variation in different regions of the face. Their findings are shown in Fig. 1.

In this paper, we propose a facial emotion recognition scheme using supervised adaptive resonance theory (ARTMAP). The proposed scheme formulates a feature vector based on the facial animation parameters (FAP) corresponding to the emotional region of the face, as shown in Fig. 1, and utilizes a default ARTMAP as a classifier for emotion recognition. The database used in this paper is Interactive Emotional Dyadic Motion Capture (IEMOCAP)

Zaheer S. and Kim J.
Facial Emotion Recognition in Presence of Speech using a Default ARTMAP Classifier.
DOI: 10.5220/0006572204360442

(Busso et al., 2008). IEMOCAP database is an acted, multimodal and multispeaker database, developed by Signal Analysis and Interpretation Laboratory (SAIL) lab at the University of Southern California (USC). It contains approximately 12 hours of audiovisual data, including audiovisuals, motion capture of face, text transcriptions. The motion capture information, the interactive setting to elicit authentic emotions, and the diversity of the actors in the data base (five males and five females) make this database a valuable, realistic and challenging emotion corpus.

This paper is organised as follows: Section 2 describes the facial feature extraction. Section 3 explains the facial emotion classification using default ARTMAP, and the classification results follow in Section 4. Finally, the concluding remarks are presented in Section 5.

2 FACIAL FEATURE EXTRACTION

Recent studies using audiovisual data similar to ours have shown that Face Animation Parameters (FAP) can be an effective feature set choice for extraction of emotional information even when the user is speaking (Kim et al., 2013; Mower et al., 2011).

> "A Face Animation Parameter (FAP) is a component of the MPEG-4 International Standard developed by the Moving Pictures Experts Group. FAP represent displacements and rotations of the feature points from the neutral face position, which is defined as: mouth closed, eyelids tangent to the iris, gaze and head orientation straight ahead, teeth touching, and tongue touching teeth" (Petajan, 2005).

2.1 Motion Capture and FAP

As FAP are distances between two points on a face, a prerequisite to calculating FAP is the availability of the motion capture data for the corresponding points on the face. Fig. 2 demonstrates the motion capture points available in the database. Combining the information from Fig. 1 and Fig. 2, desirable FAP can be calculated. Our set of 30 FAP is similar to the ones used by (Kim et al., 2013; Mower et al., 2011), with the exception of FAP corresponding the mouth openings. These FAP are shown in Fig. 3.

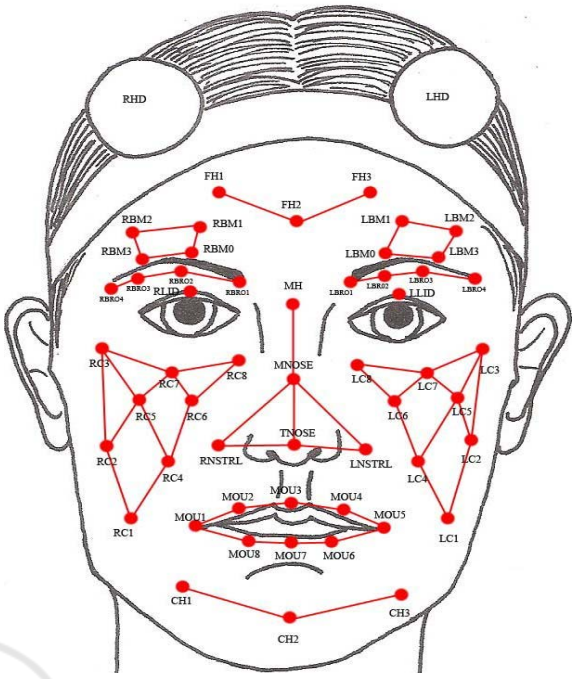

Figure 2: Visual representation of the motion capture pints on the face.

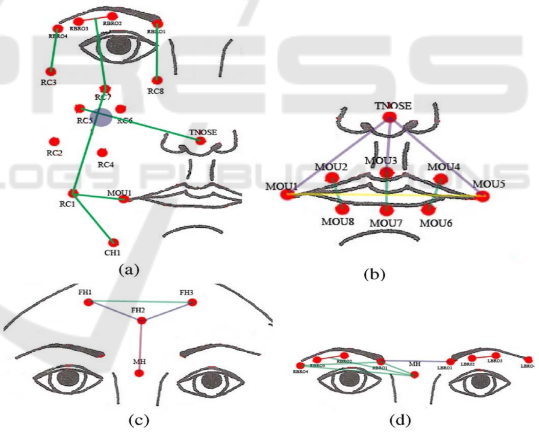

Figure 3: Visual representation of the FAP used for facial emotion recognition.

2.2 Facial Features based on FAP

After obtaining the required FAP, the next step is to generate the feature vector for facial emotion classification. Ninety (x,y,z-components of each of the FAP) FAP values are extracted from each frame of the audiovisual segment. The selected features for the audiovisual segment consists of means, standard deviations, ranges, upper and lower quartiles, and quartile ranges for all 90 values over the entire segment. Consequently, the net feature vector consists of 540 features for each audiovisual segment.

2.3 Facial Feature Normalization

The database has multiple actors and they all have distinct facial features and sizes, which means that the base values for their FAP are different. Therefore, the FAP features need to be normalized to minimize the effect of base value variation among different faces. We use z-normalization for this purpose. Mean and standard deviation for each face were calculated over the entire spectrum of emotions expressed by the corresponding actor. These mean and standard deviation values for each face are used to calculate feature values in terms of z-scores using:

$$FAPFeat_{zscores} = \frac{(FAPFeat - \mu_{FAPFeat})}{\sigma_{FAPFeat}}, \quad (1)$$

where $FAPFeat$ are the FAP-based features, $\mu_{FAPFeat}$ and $\sigma_{FAPFeat}$ are the means and standard deviations, respectively, of the features across the entire spectrum of emotions.

2.4 Facial Feature Scaling

The classifier for facial emotion recognition is a default ARTMAP neural network. Since default ARTMAP is a variant of fuzzy ARTMAP, the inputs to the network need to be scaled to a zero-to-one range. (2) is used for scaling.

$$FAPFeat_{scld} = \frac{(FAPFeat_{zscores} - FAPFeat_{min})}{(FAPFeat_{max} - FAPFeat_{min})}, \quad (2)$$

where $FAPFeat_{max}$ and $FAPFeat_{min}$ are the maximum and minimum values, respectively.

3 FACIAL EMOTION CLASSIFICATION USING ARTMAP

Even though FAP based features have been shown to be quite effective for facial emotion, there are some hindering issues in the choice of classifiers. These issues stem from the way in which the feature vectors are formulated. A common practice is to accumulate FAP over a segment or an utterance, and then formulate a feature vector by applying statistical operations over the accumulated FAP. The statistical operations applied in this case are: mean, standard deviation, range (max-min values), upper quartile, lower quartile, and quartile range. This results in relatively large feature vectors with a fewer training instances because each instance is sampled over utterances/segments containing hundreds of frames.

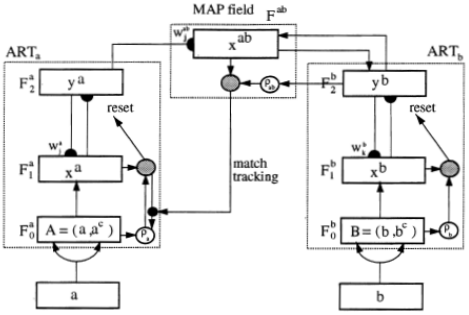

Figure 4: The architecture of Fuzzy ARTMAP.

Therefore, fewer training instances (<3000) with a feature vector size of over 500 presents a particular case of 'curse of dimensionality'. This issue has been tackled in research by coupling a classifier (Neural network or SVM) with a feature dimensionality reduction implemented through Information Gain (IG), Principal Feature Analysis (PFA) , Deep Belief Network (DBN), etc (Kim et al., 2013).

However, in this research, we opted for a Fuzzy ARTMAP classifier, Default ARTMAP classifier specifically. We chose ARTMAP because it enables fast learning by simultaneously clustering/categorizing and classifying. Additionally, ARTMAP is plastic while maintaining spasticity, i.e., it can learn new information without forgetting what it already has learnt. Before explaining the Default ARTMAP classifier that we employed in this research, the following subsection will provide some back ground on Fuzzy ARTMAP, in particular, and Fuzzy Adaptive Resonance Theory (ART).

3.1 Fuzzy ARTMAP

Fuzzy ARTMAP, or supervised ART, is a combination of two ART neural networks that are connected through a MAP field (shown in Fig. 4) (Carpenter et al., 1991b). The first Fuzzy ART neural network, ART_a, categorizes the inputs, while the second one, ART_b, categorizes the output class labels. The association between the two categorizations is mapped via a MAP field, hence the name ARTMAP.

Fuzzy ART implements fuzzy logic into ART's pattern recognition, thus enhancing generalizability (Carpenter et al., 1992). The first step in Fuzzy ART learning is complement coding. This is done by concatenating fuzzy complement of the input at the end of the input vector:

$$A = (a|a^c). \quad (3)$$

After complement coding the inputs, Fuzzy ART is initialized by categorizing the first input and initializing the weights and vigilance parameter, ρ. The vigi-

lance parameter controls the level of fuzzy similarity acceptable to be categorized into the same category node. The higher ρ means stricter categorization and hence more category nodes. Once the Fuzzy ART has been initialized, the next input is selected and the activation signals to the committed nodes:

$$T_j = |\mathbf{A} \wedge \mathbf{w}_j| + (1-\alpha)(M - |\mathbf{w}_j|). \quad (4)$$

Then, the activated nodes are checked for template matching, or resonance, using the following criterion:

$$\rho|\mathbf{A}| - |\mathbf{x}| \leq 0, \quad (5)$$

where $\mathbf{x} = \mathbf{A} \wedge \mathbf{w}_j$. If there is a match, the weights are updated using:

$$\mathbf{w}_j = (1-\beta)\mathbf{w}_j + \beta(\mathbf{A} \wedge \mathbf{w}_j), \quad (6)$$

where β is the learning rate. On the other hand, if there is no match, a new new node j associated to the input is created:

$$\mathbf{w}_j = \mathbf{A}. \quad (7)$$

In a Fuzzy ARTMAP, as well as a Default ARTMAP, the same ART categorization and learning scheme is used. However, in the supervised case, the vigilance parameter for the categorization is controlled via the labels coming through the ART_b and MAP field. Further explanation on that follows in the next subsection.

3.2 Default ARTMAP

Default ARTMAP was used as the facial emotion classifier. The default ARTMAP (Amis and Carpenter, 2007; Carpenter and Gaddam, 2010) is a fuzzy ARTMAP with distributed coding for testing. Instead of winner-takes-all (WTA) testing in the typical fuzzy ARTMAP, the default ARTMAP employs the coding field activation method (CAM) (Carpenter et al., 1991a) for distributed testing. The training process for default ARTMAP is trained as follows (Fig. 5 (Amis and Carpenter, 2007)):

1. Complement code M-dimensional training set feature vectors, a, to produce $2M$-dimensional input vectors, A
2. Select the first input vector, A, with associated actual output class, K.
3. Set initial weights.
4. Set vigilance, ρ, to its baseline value and reset the code: $y = 0$.
5. Select the next input vector A, with associated actual output class, K.
6. Calculate signals to committed coding nodes

$$T_j = |\mathbf{A} \wedge \mathbf{w}_j| + (1-\alpha)(M - |\mathbf{w}_j|) \quad (8)$$

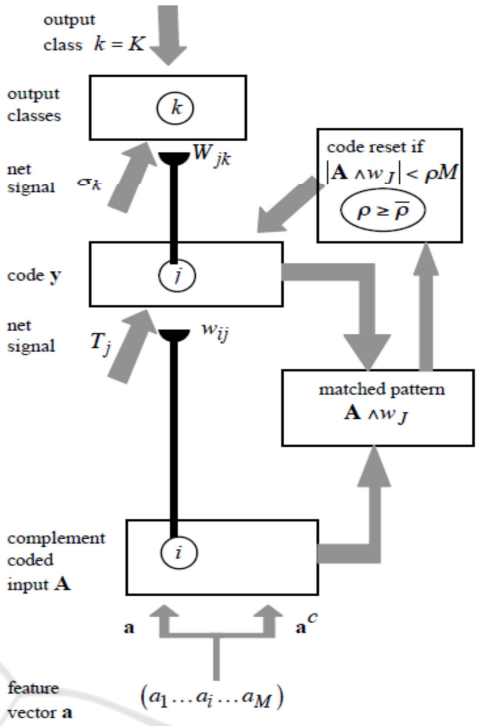

Figure 5: Default ARTMAP notation.

7. Sort the committed coding nodes, N, in descending order of T_j values.
8. Search for a coding node, J, that meets the matching criterion and predicts the correct output class, K.
9. For the next sorted node that meets matching criteria, set $y_J = 1(WTA)$
10. If the active code, J, predicts the actual output class, K. Otherwise, increase the ρ to add a new node and redo initializing and matching.
11. Update coding weights and go to 4.

After the default ARTMAP is trained, the testing is performed in the following steps:

1. Complement code M-dimensional training set feature vectors, a, to produce $2M$-dimensional input vectors, A
2. Select the first input vector.
3. Reset the code: $y = 0$.
4. Calculate signals to committed coding nodes

$$T_j = |\mathbf{A} \wedge \mathbf{w}_j| + (1-\alpha)(M - |\mathbf{w}_j|) \quad (9)$$

5. Let $\Lambda - \{\lambda - 1 \ldots C : T_\lambda > \alpha M\}$ and $\Lambda' - \{\lambda - 1 \ldots C : T_\lambda - M\} = \{\lambda - 1 \ldots C : \mathbf{w}_j = \mathbf{A}\}$.
6. Apply Increased Gradient (IG) CAM Rule to calculate y_j (Fig. 6).

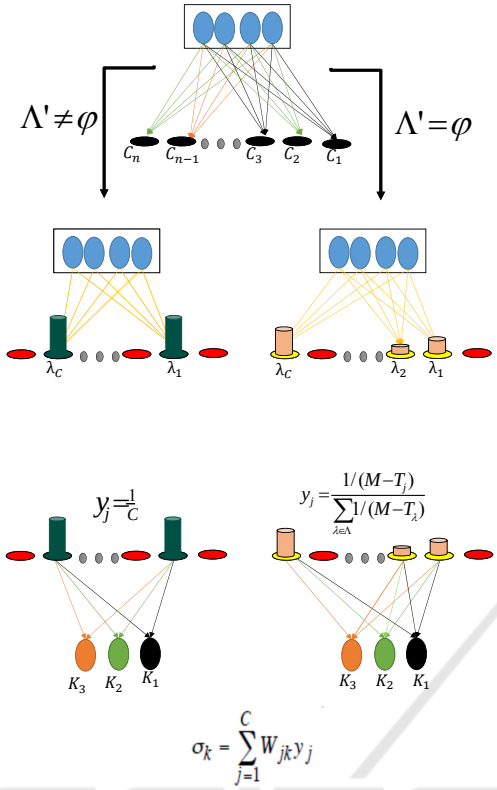

Figure 6: Default ARTMAP testing using Increased gradient CAM.

7. Calculate distributed output predictions: $\sigma - \sum_{j=1}^{C} \mathbf{W}_{jk} y_j$
8. Predict output classes from σ_k

3.3 Facial Emotion Recognition using default ARTMAP

Default ARTMAP was used as the classifier for Facial emotion recognition. The configuration of the ARTMAP network is shown in Fig. 7.

The Default ARTMAP was trained using four-fold cross validation using 2442 training instances from eight actors/speakers (4 males and 4 females). The followings are the configuration parameters used:

- Learning rate: 0.7
- Choice parameter, α: 0.27
- Base vigilance: 0.2
- CAM rule parameters: 1

After Default ARTMAP for facial parameters was trained, we got a facial emotion ARTMAP classifier with the following configuration: 1080 input nodes, 1112 category nodes, and five class nodes corresponding to five emotion classes: neutral, happiness, anger, frustration, and sadness.

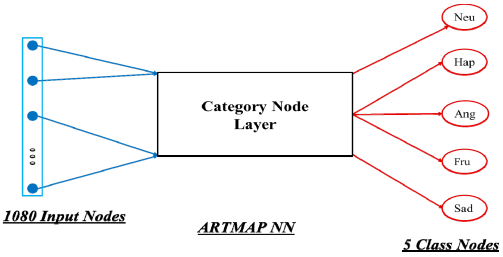

Figure 7: Configuration of the trained ARTMAP classifier for Facial emotion recognition.

Table 1: Confusion matrix for the FAP-based facial emotion classifier.

	Neutral	Happy	Angry	Frust.	Sad
Neutral	0.529	0.059	0.059	0.235	0.118
Happy	0.019	0.830	0.057	0.075	0.019
Angry	0.036	0.071	0.5	0.393	0
Frust.	0.024	0.072	0.241	0.590	0.072
Sad	0.027	0.108	0.054	0.243	0.568

4 TESTS AND RESULTS

After training, the classifier was tested using the training data from the remaining two actors (one male, one female) in the IEMOCAP database. In other words, the ARTMAP was trained using eight of the 10 actors in IEMOCAP and tested using the other two. The classification results showed a five class classification accuracy of over 68%. The confusion matrix for the Default ARTMAP body language classifier is shown in Table 1. As evident from the confusion matrix, the most frequent instances of misclassification/confusion occurred between angry and frustrated. This confusion is understandable as these two emotions are often not easily distinguishable even for humans.

We also compared our results with existing similar researches on IEMOCAP. These approaches used support vector machines (SVM) preceded by feature dimension reduction. Table 2 shows comparative results of default ARTMAP against the following:

- SVM with Reynolds Boltzman Machine (RBM-SVM) (Shah et al., 2014)
- SVM with Principal Feature Analysis (PFA-SVM) (Kim et al., 2013)
- SVM with Deep Belief Networks (DBN-SVM) (Kim et al., 2013)
- Emotion profiled SVM (EP-SVM), where each one-vs-all emotion classifiers used a feature vec-

Table 2: Comparative Results on IEMOCAP using FAP-based features.

Classification approach	Accuracy
RBM-SVM (Shah et al., 2014)	60.71%
PFA-SVM (Kim et al., 2013)	65%
DBN-SVM (Kim et al., 2013)	68%
EP-SVM (Mower et al., 2011)	71%
Default ARTMAP	72.2%

tor profiled for that particular emotion (Mower et al., 2011)

These results are for four class (neutral, happy, anger, as sadness) classification as those researches used four class classification. It is evident from the table that our approach gives the best results for FAP-based classifier on IEMOCAP data set. Furthermore, (Mower et al., 2011) and (Kim et al., 2013) used both facial and vocal features. However, since they had a similar set of facial features and they tested their approaches on IEMOCAP, we used their results for comparison as well.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed facial emotion recognition using a default ARTMAP classifier. The proposed classification scheme along with the FAP-based features was shown to be an effective facial emotion classifier in the presence of speech. The results show that our approach also yielded better results than the existing state-of-the-art on IEMOCAP database.

In future, we plan to integrate our emotion recognition with real time perception. Furthermore, we also intend to investigate other configurations of ARTMAP involving distributed training along with the distributed testing used in this paper.

ACKNOWLEDGEMENT

This work was supported by the ICT R&D program of MSIP/IITP. [2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion]

REFERENCES

Amis, G. P. and Carpenter, G. A. (2007). Default artmap 2. In *2007 International Joint Conference on Neural Networks*, pages 777–782.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335.

Carpenter, G. A. and Gaddam, S. C. (2010). Biased art: A neural architecture that shifts attention toward previously disregarded features following an incorrect prediction. *Neural Networks*, 23(3):435 – 451.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., and Rosen, D. B. (1992). Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. *Trans. Neur. Netw.*, 3(5):698–713.

Carpenter, G. A., Grossberg, S., and Reynolds, J. H. (1991a). Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, 4(5):565 – 588.

Carpenter, G. A., Grossberg, S., and Rosen, D. B. (1991b). Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4(6):759 – 771.

Hirota, K. and Dong, F. (2008). Development of mascot robot system in nedo project. In *Intelligent Systems, 2008. IS '08. 4th International IEEE Conference*, volume 1, pages 1–38–1–44.

Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audio-visual emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Li, H., Lin, Z., Shen, X., Brandt, J., and Hua, G. (2015a). A convolutional neural network cascade for face detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, W., Li, M., Su, Z., and Zhu, Z. (2015b). A deep-learning approach to facial expression recognition with candid images. In *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*, pages 279–282.

Liu, P., Han, S., Meng, Z., and Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Z.-T., Min, W., Dan-Yun, L., Lue-Feng, C., Fang-Yan, D., Yoichi, Y., and Kaoru, H. (2013). Communication atmosphere in humans and robots interaction based on the concept of fuzzy atmosfield generated by emotional states of humans and robots. *Journal of Automation, Mobile Robotics and Intelligent Systems*, 7(2):52–63.

Mariooryad, S. and Busso, C. (2016). Facial expression recognition in the presence of speech using blind lexical compensation. *IEEE Transactions on Affective Computing*, 7(4):346–359.

Mower, E., Mataric, M. J., and Narayanan, S. (2011). A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070.

Petajan, E. (2005). *MPEG-4 Face and Body Animation Coding Applied to HCI*, pages 249–268. Springer US, Boston, MA.

Rozgi, V., Ananthakrishnan, S., Saleem, S., Kumar, R., and Prasad, R. (2012). Ensemble of svm trees for multimodal emotion recognition. In *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4.

Shah, M., Chakrabarti, C., and Spanias, A. (2014). A multimodal approach to emotion recognition using undirected topic models. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 754–757.