The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Context-Centric Approach in Paralinguistic Affect Recognition System

Andreas Marpaung,[1,] **Avelino Gonzalez**[2]
[1, 2] University of Central Florida, Orlando, FL
[1]amarpaung@knights.ucf.edu, [2]avelino.gonzalez@ucf.edu

Abstract

As the field of paralinguistic affect recognition has become more mature, many researchers have shifted their approach from a single channel of affect manifestation to a multi-modal one in developing their affect recognition systems. In the spirit of continuing this trend in multi-modal work, our work utilizes paralinguistic features of speech and contextual knowledge. Through our human study, we found that contextual knowledge had positive impact on a human's affect recognition ability when combined with paralinguistic features of speech. In this research, we propose a novel architecture called Context-Based Paralinguistic Affect Recognition System (CxBPARS) that combines the traditional paralinguistic affect recognition approach using classification algorithms and the contextual knowledge related to the emotion elicitors and their environment. By combining the results of an AdaBoost classifier and contextual modeling, we achieved an improvement in affect recognition accuracy from 29.5% (context free) to 53.0% (context dependent).

Introduction

Humans have difficulty interpreting the expressed emotions of other people when not in a face-to-face interaction. This can be alleviated when the context of conversation is known (Calvo and D'Mello 2010). For instance, one may find it hard to understand an emotionally-charged person with a heavy foreign accent through a phone line. However, when we know that the conversation occurs in a call center context, we may be able to associate the high-pitch tone with anger or frustration resulting from a complaint. A misinterpretation may also occur when person A hears a foreign speaker (person B) speaking in a language completely unknown to A. Because A cannot understand any words semantically, A can interpret a high-pitch tone by B to be a happy expression when it is expressed on a festive occasion.

There is great merit in advancing the state-of-art in paralinguistic affect recognition. The current trend has shifted from the unimodal affect recognition approach to a multi-modal one as the multi-modal has produced better results than the unimodal approach. The implementation of deep learning algorithms has also gained some popularity in the affect recognition research. Some of the latest multi-modal works include the combination of facial expression, text, and acoustic features (Sadoughi and Busso 2019), and the combination of visual and audio signals (Ren et al 2019). Another multi-modal work utilizing deep learning algorithms is that of Zhang et al. (2016, 2018); these works utilized Convolutional Neural Networks (CNN) and 3D-CNN to produce audio-visual segment features. It fused these audio-visual segment features in a Deep Belief Networks (DBN). Huang et al. (2019) combined both linguistic and paralinguistic features with CNN to recognize affect. This multi-modal trend is also reflected through the entries to many affect recognition competitions for the past decade. Some of these competitions include the Audio/Visual Emotion Challenge (AVEC) (Ringeval et al. 2019), the Emotion Recognition in the Wild Challenge (EmotiW) (Dhall et al. 2019), and the International Speech Communication Association (ISCA) INTERSPEECH COMputational PARalinguistic challengE (COMPARE) competition (Schuller et al. 2019). With this trend in mind, we want to add a new different mode—contextual knowledge.

Modeling context is not easy, especially when it involves open-ended cases and is not domain specific. Some research work in Context- Based Reasoning (CxBR) (Gonzalez et al. 2008) and Context Mediated Behavior (CMB) (Turner 1998) can model context but they are limited to narrow and pre-determined scenarios, thus making them impractical for our purposes. To our knowledge, very few works relate context to affect recognition. Hammal and Suarez (2013) pioneered context-based affect recognition workshops where the investigators explored the effect of contextual information that can provide different nuances and complexities in developing human-centric systems for affect recognition but none of the works focused on paralinguistic speech. We investigated several ways to utilize the contextual information that led to the conception of our proposed architecture. To our knowledge, no reported research integrates a

non-domain-specific contextual modeling approach with the traditional paralinguistic affect recognition.

For an apples-to-apples comparison, we chose to compare our work with Gosztolya et al.'s work (2013) - the entry to the Emotion sub-challenge of INTERSPEECH 2013 (Schuller et al., 2013) that had the highest classification accuracy results among all entries. A more inclusive comparison between our approach and all entries to the Emotion sub-challenge can be found in Marpaung (2019). INTERSPEECH conferences have hosted the Computational Paralinguistic Emotion (ComParE) Challenges intermittently since 2009. ComParE 2013 focused on four sub-challenges: (a) social signal challenge, (b) conflict challenge, (c) emotion challenge, and (d) autism challenge. We were interested in ComParE 2013 because it was the only one that dealt with affect recognition that utilized the GEMEP corpus (Banziger et al. 2011). To our knowledge, GEMEP corpus is the only corpus that has documented different scenarios used to guide the actors to enact certain emotions during the recording process.

First, we describe our previous human study effort to better understand the influences of contextual knowledge in a human's affect recognition ability. After describing our human study, we give some background information on the emotion lexicon and sentiment analysis domain. The last several sections focus on our architecture and algorithm, our experimental results, conclusion, and future work plan. For the record, our research focuses on 17 emotions: admiration, amusement, anger, anxiety, contempt, despair, disgust, fear, interest, irritation, joy, pleasure, pride, relief, sad, surprise, and tenderness.

Human Study

As a prelude to our research in context-based computational paralinguistics, we studied the impact of contextual knowledge on a human's affect recognition. For this study, we utilized the GEMEP corpus. For details of this study, we refer interested readers to (Marpaung and Gonzalez 2017). In summary, we manually extracted two pieces of contextual information from the description of the situation as provided by GEMEP: (1) action context, and (2) relationship context. The action context is defined as any information about what happens to a person or a group of people involved in the conversation; the relationship context is defined as any information that relates an object or an environmental element to a person or relates a person to another person (or to a group of people). For each audio file, each test subject went through a three-phase process: (1) listened to the audio file only, (2) listened to the same audio file after being given one piece of contextual information (either the action or the relationship context), and (3) listened to the same audio file after being given a second piece of contextual information. In each phase, the participants selected the emotion that they believed that the speaker in sound bite was expressing, and rate their confidence level on their selected emotion. To further understand the effect of action and relationship contextual knowledge, we exposed both contextual knowledge in two different sequences of presentation: (1) no context – action context only – action and relationship contexts; and (2) no context – relationship context – relationship and action contexts.

Our study concluded that there was significant positive impact in adding contextual knowledge to a human's affect recognition ability through the paralinguistic features of speech. Between the two identified contextual information, we found that the action context had a more significant impact on the correctness and confidence level in human affect recognition ability than did the relationship context.

Sentiment Analysis: Emotion Lexicon

Inspired by an increasing attention to the field of Sentiment Analysis, we borrowed the emotion lexicon, a concept from the Sentiment Analysis research. Sentiment Analysis (SA) operates at the intersection of information retrieval, natural language processing, and artificial intelligence. SA focuses on determining the sentiment expressed in text. This field of research studies the sentiment, the phenomena of opinion, evaluation, appraisal, attitude, and emotion (Liu 2012)

Mohammad (2016) reviews some competitions on valence classification that include: (1) Sentiment Analysis in Twitter (SAT), held in 2013, 2014, and 2015, (2) Aspect Based Sentiment Analysis (ABSA), held in 2014 and 2015, and (3) Sentiment Analysis of Figurative Language in Twitter and Sentiment Analysis on Movie Reviews, held in 2015. Poria et al. (2016) proposed a multimodal sentiment analysis framework that fused relevant features for text and visual data. McDuff et al. (2015) found that voter preference (or opinion) could be determined with an accuracy of 73% through facial expressions in videos by analyzing 611 responses to five video clips of a US presidential election debate.

Word and Affect Mappings

To build the contextual knowledge necessary for our work, we utilized the emotion lexicon that correlates many English words to valence and arousal. Our work utilized the emotion lexicon by Warriner, Kuperman, and Brysbaert (2013), which contains affective words with 13,915 English lemmas (63.5% nouns, 12.6% verbs, 22.5% adjectives, and 1.4% unspecified parts of speech). We utilize this dictionary because it is not a domain-specific. This emotion lexicon also provides the largest collection of words that covers primary demographic information, such as age, gender, and education. These words were mapped in the semantic space (SS) with

three axes: valence, arousal, and dominance. Valence measures how pleasant the stimulus is, with the scale ranges from 1 (unpleasant) to 9 (pleasant). Arousal measures how high/low the intensity of emotion (excitement) provoked by the stimulus is, with the scale ranging from 1 (calm) to 9 (excited). Dominance measures how much control is exerted by the stimulus, with the scale ranging from 1 (in control) to 9 (controlled). Because most emotion theories in Psychology correlate affect with valence and arousal only, our work uses only these two. We also mapped the entire 17 emotions to the semantic space and referred them as Affect Vectors (AVs). We utilized the same keywords used in the GEMEP corpus for the AVs. For example, we used the keyword *fear* instead of *fearful* or *scare*.

Word - Affect Relationship

This section focuses on how we map the keywords, obtained from the scenario description and the context database. By mapping the emotions and the keywords, we can calculate the distances between each keyword to the entire 17 emotions and associate the keyword of interest to the affect with the shortest distance.

Similar to AVs, the keywords are also mapped into the SS in the same fashion. The Valence Coordinate (VC) and Arousal Coordinate (AC) values for each keyword vector are obtained from Warriner et al.'s dictionary (2013). Figure 1 shows the words *lottery* (VC = 6.33, AC = 6.05), *win* (VC = 6.97, AC = 5.61), *receive* (VC = 7.14, AC = 4.3), and *gift* (VC = 7.27, AC = 4.64) in SS.

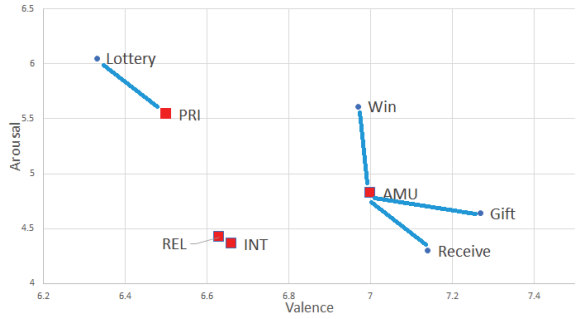

Figure 1: Word - Affect Relationship

To estimate the closest emotion associated to each keyword (from the 17 emotions), the Euclidian distance is measured between each keyword and the affect coordinates. For example, the calculation of Euclidian distances between keyword *lottery* to some of the emotions are: (1) to pride: 0.54; (2) to relief: 1.66; (3) to interest: 1.72; and (4) to amusement: 1.40. Once calculated, the emotion with the shortest distance to the keyword of interest is associated with this keyword. Based on this calculation, the keyword lottery is best associated to pride. Refer to Figure 1 for their graphical representations.

Architecture

Supported by our previous study on human subjects (Marpaung and Gonzalez 2017), we propose the CxBPARS system and its novel architecture. Figure 2 shows the overall schematic of CxBPARS.

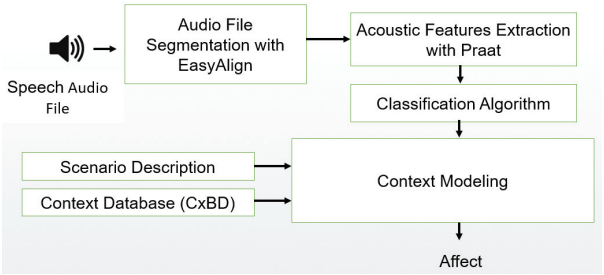

Figure 2: CxBPARS Architecture

Our approach focuses on extracting keywords from the description of the situation, and then evaluating the emotion implications caused by these words. The emotion implications are then used to disambiguate classifier results.

Segmentation and Features Extraction

The audio files from the GEMEP corpus were segmented manually at the word-level using EasyAlign (Goldman 2011). From these segmented audio files, five acoustic features were extracted using Praat (Boersma 2001). EasyAlign is a freely-distributed system usable as a plug-in to the Praat software. The five features were: (1) mean pitch—the average of the sound's pitch, which shows the highness or lowness of the human voice, measured in Hertz, (2) jitter—pitch's variations in human's voice, which causes a rough sound, (3) shimmer—a frequent back and forth change in amplitude from soft to louder, (4) mean harmonics-to-noise ratio (HNR)—the ratio between multiples of fundamental frequency and the noise, and (5) mean noise-to-harmonics ratio (NHR)—the ratio between multiples of the noise and fundamental frequency (the inverse of HNR).

From the GEMEP corpus, 579 frames were segmented. The frame distributions were as follows: admiration (18), amusement (37), anger (39), anxiety (40), contempt (23), despair (42), disgust (22), fear (32), interest (44), irritation (45), joy (39), pleasure (43), pride (35), relief (37), sad (40), surprise (18), and tenderness (24). From these frames, the five acoustic features were extracted. This process produced 2,895 (579 x 5) acoustic features.

Classification Algorithm

This research utilized an AdaBoost classifier implemented in Weka (Waikato Environment for Knowledge Analysis) (Frank et al 2016) using a 10-fold cross validation process. Weka implemented AdaBoost M1 method based on Freund and Schapire (1996).

We chose this classifier because we were interested in directly comparing our approach with Gosztolya et al.'s research (2013), as it had the highest classification accuracy in the INTERSPEECH 2013 Emotion Sub-Challenge. The input of the classifier was the 2,895 acoustic features extracted by Praat. The output of the classifier is a tuple of the most likely emotion the speaker conveys (A) and its probability (p): (A_n, p_n). Since we focus on 17 different emotions, each classifier produces 17 output tuples: {(A_1, p_1), (A_2, p_2) … (A_{17}, p_{17})}.

Scenario Description

The GEMEP corpus utilized the felt enacting technique during the recording process. This technique guided the actors to use personal experiences to enact certain emotions based on the given scripts originally written in French. To build our context-aware system, these scenario scripts were translated from French to English using Google Translate (2019). Although the GEMEP scenario descriptions are long and complex, these scenarios roughly represent someone's daily activities, and can equate to the cryptic descriptions someone might enter into their smart phone calendar or notes applications. To clarify, no two-way verbal conversations occurred in any of these scenario descriptions.

For our initial approach, two action keywords were chosen manually from the translated scenarios, and their VC and AC values were mapped to the semantic space. The VC and AC values were obtained from the emotion lexicon by Warriner, Kuperman, and Brysbaert (2013). We extracted the initial keywords manually and randomly without knowing their positions in the semantic space.

Context Database (CxDB)

For this work, we introduce the Context Database (CxDB) that mimics a human's prior knowledge. Each piece of contextual knowledge in the CxDB has these attributes: (1) context name, (2) keywords associated with the contextual knowledge, and (3) possible emotions that can be triggered within that context. An example of contextual knowledge is shown below.

Context Name: Meet_Important_Person
Keywords: admire, respect, accomplish, inspiration, famous, recognized
Applicable emotions: admiration, amusement, anxiety, fear, interest, joy, pleasure, pride.

As our initial approach, we built the CxDB manually with 50 contexts containing the information above. We acknowledge that building this database by hand as we did is not the ultimate solution, but we first wanted to determine whether this approach worked before embarking on research to automate the creation of this database.

Context Modeling

The proposed algorithm for our framework is as follows.

1. Execute the AdaBoost classifier using the 10-fold cross validation process (input: segmented audio file; output: classification result (R_i) with the format of the affect (A_i) and its probability (p_i): (p_1, p_2… p_{17}) where i = 1 to 17, p is 0.0 to 1.0, and R_i = (A_i, p_i).
2. Rearrange the results in decreasing order from the highest to the lowest probability.
3. Pick the top X results (those with the highest X probabilities). X represents the search area for the context modeling; X is selected empirically. For this research, we picked X = 10, X = 11, X = 12, and X = 17. Based on our experiments, X = 11 gave the best result while any numbers below or above this number produced poorer results. We ran X = 17 as the inclusivity case.
4. Hand-pick two keywords (SD1 and SD2) from the scenario related to the audio file. We focused on selecting the verbs/action words. Each keyword has VC and AC values based on the Emotion Lexicon.
5. For each entry in CxDB, compare all keywords (Keyword_1, Keyword_2, …., Keyword_K) in each context (Context_1, Context_2, …, Context_50) with the scenario description keywords (SD1 and SD2). Pick the context with the highest number of (lexically) matched words and call it CxDBwinner. If there is a tie between two contexts, simply pick the one occurring first.
6. From CxDBwinner, extract all applicable emotions. We called them *Matched Contexts*.
7. Look for some agreements between emotions in *Matched Contexts* and the *Classifier Results* (A_{top_1}, A_{top_2}, …, A_{top_n}) by finding the intersection between these two sets.
8. Extract the VC and AC for all affects in A_{top_x}, …., A_{top_y} from the Emotion Lexicon by Warriner, Kuperman, and Brysbaert (2013).
9. Calculate the Euclidean distance using the VCs and ACs between the keywords (SD1 and SD2) and each emotion in (A_{top_x}, …., A_{top_y}).
10. The affect with the shortest distance is declared as the winner.
11. STOP.

Next, we describe our experimental results to test our approach.

Experimental Results

We divided our experiments into three phases: (1) Phase 1—classifier, (2) Phase 2—context-centric approach, and (3) Phase 3—context-centric approach with intelligently

selected keywords. For Phase 1, we ran our classifier with: (a) 12 emotions, as we wanted to compare our approach with Gosztolya et al.'s previous research, and (b) 17 emotions, as we sought to advance the research in CPAR domain by working with more subtle emotions. For Phase 2, we utilized our context modeling approach in 17 emotions to break new ground in the state-of-the-art. For Phase 3, we intelligently selected the keywords (SD1 and SD2) to study the keyword's sensitivity and its influence on the architecture's classification results.

Phase 1: Classifier

For Phase 1, we utilized an AdaBoost classifier implemented in Weka. Table 1 compares our UAR (Unweighted Average Recall) results for the 12 and 17 emotions with Gosztolya et al's result. At the face value, our result was lower by 2.8 percentage points. Refer to Marpaung (2019) for the comparison between our results and the winners of the INTERSPEECH 2013 Emotion sub-challenge.

Next, we describe our Phase 2 experiment where we integrated the contextual knowledge to the classifier results.

	Our Results	Gosztolya et al. (2013)
12 affects	39.5%	42.3%
17 affects	29.5%	

Table 1: Comparing our results and Gostolya et al.'s result

Phase 2: Context-Centric Approach

Table 2 shows our results utilizing our CxBPARS architecture for 17 emotions for different values of X (10, 11, 12, and 17). In Phase 2, the verb/action keywords were randomly chosen without knowing their VCs and ACs with respect to the VCs and ACs of the intended emotions.

	Our Results
X = 10	52.8 %
X = 11	53.0 %
X = 12	50.3 %
X = 17	38.7 %

Table 2: CxBPARS Experimental Results

As shown in Table 2, the architecture gave the best classification result when we set the value of X to 11.

Phase 3: Keywords Sensitivity

From Phase 2, we chose X=11 and conducted another study of Phase 3 to measure the sensitivity of the keyword selection. In this phase, we changed one of the two keywords for each emotion (amusement, anxiety, and relief) and ran the algorithm again using the 10-fold cross validation process. We chose the keywords whose VCs and ACs close to the intended emotions on purpose. Table 3 shows the results of Phase 3.

For the scenario whose ground truth was amusement, the original keywords were *figure* and *laugh*. For Phase 3, we changed the keyword *figure* to *create*. For Phase 2, we achieved 22 correct cases (out of 37) and for Phase 3, we reached 36 correct cases. For anxiety, our result improved from 4 correct cases to 39 cases (out of 40) when we swapped the keyword *trouble* with *distress*. And finally, for relief, we reached 37 correct cases (out of 37) when we switched *suspect* with *arise*.

Overall, using this intelligently selected keywords method in Phase 3, our results improved from 53.0% (in Phase 2) to 58.4% (in Phase 3). The experiment therefore indicated that the choice of keywords does influence the accuracy results. If a higher accuracy result, the valence and arousal coordinates must be located near to the coordinates of the emotions. Otherwise, the accuracy results are negatively impacted.

Emotions	Phase 2	Phase 3
Amusement		
Keywords	Figure (VC = 5.0, AC = 3.67) Laugh (VC = 8.05, AC = 5.39)	Create (VC = 7.06, AC = 4.86) Laugh (VC = 8.05, AC = 5.39)
Accuracy	22 (out of 37)	36 (out of 37)
Anxiety		
Keywords	Trouble (VC = 2.87, AC = 5.6) Fail (VC = 2.33, AC = 5.5)	Distress (VC = 2.37, AC = 4.5) Fail (VC = 2.33, AC = 5.5)
Accuracy	4 (out of 40)	39 (out of 40)
Relief		
Keywords	Suspect (VC = 2.39, AC = 4.6) Favor (VC = 6.67, AC = 4.61)	Arise (VC = 6.61, AC = 4.49) Favor (VC = 6.67, AC = 4.61)
Accuracy	4 (out of 37)	37 (out of 37)

Table 3: Phase 3 Results

Conclusion and Future Works

Inspired by our human study, we proposed the CxBPARS architecture where we employed both the paralinguistic speech features and the relevant contextual knowledge. Through our experiments, we could see the improvement from 29.5% (context free in Phase 1) to 53.0% (context dependent in Phase 2). Utilizing the intelligently selected keywords method, our accuracy improved from 53.0% (in Phase 2) to 58.4% (in Phase 3). We realized that this proof-of-concept work was not automated. Thus, these results should be considered as an upper bound.

Our future works include the automation of creating and maintaining the CxDB and the automation of the intelligent

keyword selection process by utilizing the information extracted from the smartphone applications.

References

Ba¨nziger, T.; and Scherer, K.R. 2011. Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus." *In K. R. Scherer, T. Ba¨nziger, & E. B. Roesch (Eds.), Blueprint for affective computing: A sourcebook* (pp. 271–294). Oxford, England: Oxford University Press.

Boersma, P.; and Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.37 retrieved 14 March 2018 from http://www.praat.org

Calvo, R.A.; and D'Mello, S.K. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and their Applications. *IEEE Transactions on Affective Computing,* 1(1), 18-37.

Dhall, A.; Goecke, R.; Ghosh, S.; and Gedeon, T. 2019. EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. In *ACM International Conference on Mutimodal Interaction*.

Eugenio, M.; Teresa, M.; and Urena-Lopez, A. 2014. Sentiment analysis in Twitter. Vol. 20. Issue 1 (Jan 2014) pp. 1-28 https://doi.org/10.1017/S1351324912000332

Frank, E.; Hall, M.; and Witten, I. 2016. The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Fourth Edition.

Freund, Y.; and Schapire, R. 1996. Experiments with a new boosting algorithm, Machine Learning: *Proceedings of the Thirteenth International Conference,* 148–156.

Goldman, J. 2011. EasyAlign: an automatic phonetic alignment tool under Praat," *In: Interspeech'11, 12th Annual Conference of the International Speech Communication Association*. Firenze (Italy).

Gonzalez, A.J.; Stensrud, B.S.; and Barrett, G. 2008. Formalizing context based reasoning: A modeling paradigm for representing tactical human behavior. *Int. J. Intell. Syst*. Vol. 23. pp. 822–847

Google, Inc. https://www.google.com/, (2019).

Gosztolya, G.; Róbert, B.; and László, T. 2013. Detecting autism, emotions and social signals using adaboost, *In INTERSPEECH 2013*, pp. 220-224.

Hammal, Z.; and Suarez, M.T. 2015. Towards context based affective computing introduction to the third international CBAR 2015 workshop. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG),* Ljubljana. pp. 1-2. doi: 10.1109/FG.2015.7284841.

Huang, K.-Y.; Wu, C.-H.; Hong, Q.-B.; Su, M.-H.; and Chen, Y.-H. 2019. "Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds." In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5866–5870.

Liu, B.; and Zhang, L. 2012. A survey of opinion mining and sentiment analysis. *In Aggarwal, C. C., & Zhai, C. (Eds.), Mining Text Data,* pp. 415–463. Springer US.

Marpaung, A.; and Gonzalez, A. 2014. Toward building automatic affect recognition machine using acoustics features. *In: FLAIRS Conference.*

Marpaung, A.; Gonzalez, A. 2017. Can an Affect-Sensitive System Afford to Be Context Independent?". *In: Brézillon P., Turner R., Penco C. (eds) Modeling and Using Context. CONTEXT 2017. Lecture Notes in Computer Science*, vol 10257. Springer.

Marpaung, A. 2019. Context-Centric Affect Recognition From Paralinguistics Features of Speech. Ph.D. Dissertation. University of Central Florida.

McDuff, D.; Kaliouby, R.E.; Cohn, J.F.; and Picard, R.W. 2014. Predicting ad liking and purchase intent: large-scale analysis of facial responses to ads. *IEEE Trans. Affect. Comput.* Vol. 6 (3) pp. 223–235. http://dx.doi.org/10.1109/TAFFC.2014. 2384198.

Mohammad, S. 2016. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. Emotion Measurement. 2016, pp. 201-237. (2016).

Pang, B.; and Lee, L. 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends® *In Information Retrieval*: Vol. 2: No. 1–2. pp 1-135. http://dx.doi.org/10.1561/1500000011.

Poria, S.; Cambria, E.; Howard, N.; Huang, G.-B.; Hussain, A. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing Vol. 174 pp. 50–59. http://dx.doi.org/10.1016/j.neucom.2015.01. 095.

Ren, M.; Nie, W.; Liu, A.; Su, Y.; 2019. In *Visual Informatics* Vol. 3, Issue 3., September 2019, pp. 150-155. Doi: https://doi.org/10.1016/j.visinf.2019.10.003.

Ringeval, F.; Schuller, B.; Valstar, M.; Cummins, N.; Cowie, R.; Tavabi, L.; Schmitt, M.; et al. 2019. "AVEC 2019 Workshop and Challenge: State-of-Mind, Depression with AI, and Cross-Cultural Affect Recognition." In *Proceedings of the 9th International Workshop on Audio/Visual Emotion Challenge, AVEC'19.*

Sadoughi, N.; Busso, C.; 2019. Speech-driven animation with meaningful behaviors. In Speech Communication Vol. 110 (2019), pp. 90–100.

Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E.; Mortillaro, M.; Salamin, H.; Polychroniou, A.; Valente, F.; and Kim, S. 2013. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. *In Proc. of INTERSPEECH.* Lyon, France: ISCA, 2013, pp. 148–152.

Schuller, B.; Batliner, A.; Bergler, C.; Pokorny, F.; Krajewski, J.; Cychosz, M.; Vollmann, R.; Roelen, S.; Schnieder, S.; and Bergelson, E. 2019. "The INTERSPEECH 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity." In *Proc. INTERSPEECH 2019*.

Turner, R.M. 1998. Context-mediated behavior for AI applications. *In: Mira J., del Pobil A.P., Ali M. (eds) Methodology and Tools in Knowledge-Based Systems. IEA/AIE 1998. Lecture Notes in Computer Science (Lectures Notes in Artificial Intelligence)*, vol 1415. Springer, Berlin, Heidelberg

Warriner, A.B.; Kuperman, V.; and Brysbaert, M. 2013. *Behav Res* 45: 1191. https://doi.org/10.3758/s13428-012-0314-x

Zhang, S.; Zhang, S.; Huang, T.; and Gao, W. 2016. "Multimodal deep convolutional neural network for audio-visual emotion recognition." in *Proc. 6th ACM Int. Conf. Multimedia Retr. (ICMR)*, New York, NY, USA pp. 281–284.

Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; and Tian, Q. 2018. "Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition," in *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 28, no. 10, pp. 3030-3043, Oct. (2018). doi: 10.1109/TCSVT.2017.2719043.