# Multimodal Cross-context Recognition of Negative Interactions

Lefter, Iulia; Rothkrantz, Leon

**DOI**
10.1109/ACIIW.2017.8272586

**Publication date**
2017

**Document Version**
Accepted author manuscript

**Published in**
Proceedings of Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) 2017

**Citation (APA)**
Lefter, I., & Rothkrantz, L. (2017). Multimodal Cross-context Recognition of Negative Interactions. In *Proceedings of Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) 2017* (Vol. 7, pp. 56-61). IEEE. https://doi.org/10.1109/ACIIW.2017.8272586

# Multimodal Cross-context Recognition of Negative Interactions

Iulia Lefter[*][†] and Leon J.M. Rothkrantz [†][‡]
[*]*Systems Engineering Group, Faculty of Technology, Policy and Management , Delft University of Technology*
[†]*Interactive Intelligence Group, Faculty of Electrical Engineering, Mathematics and Computer Science,*
*Delft University of Technology*
[‡]*Department of Transport Telematics, Faculty of Transportation Sciences, Czech Technical University in Prague*

*Abstract*—**Negative emotions and stress can impact human-human interactions and eventually lead to aggression. From the perspective of surveillance systems, it is of high importance to recognize as soon as an interaction escalates and human intervention is needed. One of the limitations of deploying a system in real life is that in practice it can only be trained on a limited number of situations. In this paper we examined the generalization capabilities of a trained system given context change. For this purpose we developed scenarios and made audio-visual recordings in four different contexts in which negative interactions might occur. To obtain a quantification of cross-context performance we kept the test context fixed and performed training on itself (cross-validation) and on all the other contexts. To explore whether multiple examples in the training set are beneficial, we also trained the classifier on a merged corpus of the three contexts that were not used for testing. These experiments were done with audio features, video features and audio-visual feature level fusion to investigate which modality generalizes best. We found that context change generates a decrease in performance that is varying with within-contexts similarities. Merging multiple contexts for training in most cases results in performance just below the best predictive single context. Audio is the most robust modality and in most cases the performance of audio-visual fusion is very close to the one of audio.**

## 1. Introduction

Taking on the perspective of surveillance systems, this paper focuses on automatically recognizing negative interactions. The development of aggressive conduct often follows a typical escalation path, starting with negative emotions and stress, followed by overt manifestation which have specific verbal and nonverbal behavior characteristics, and can ultimately lead to violence [2], [28]. From a surveillance perspective, it is of major importance to timely detect such negative interactions and offer support such that the situation will deescalate.

Detection of negative interactions has a variety of applications, including monitoring human-human interactions in public service such as service desks [17], call-centers [21], health care, monitoring conflicts in meetings [15] as well as general public surveillance [32]. Another example is virtual reality therapy systems for anger management that give automatic feedback to patients based on their behavior.

One of the challenges of applying trained monitoring systems in real-life situations is their limited generalization capability. In practice a system can only be trained on a limited set of situations. However, human behavior is extremely complex and diverse, which leads to differences between the training material and the real-life scenes to be monitored. This is supported by studies in speech emotion recognition that indicate a significant inferiority in cross-compared to intra-corpus recognition accuracies [26], [19], [29], [31], while in [8] the authors studied cross-corpus video-based action detection.

Context is a key feature in behavior interpretation [24] and the research community is making more efforts towards context-sensitive systems [11]. Furthermore, context change is a possible source of limited generalization: in different contexts, different behaviors are to be expected. For example, negative interactions can appear between customers and employees at a service desk or at a vending machine. Each context can be characterized by several traits, such as expected behaviors, likely sources of conflict, length of the interaction, expected movements, number of people in the scene. For the audio analysis, there can be differences in room acoustics, noise levels and language. For video, differences in viewing angles, occlusions, and lighting conditions. Nevertheless, for both audio and video, probably one of the most important challenges is how to handle the wide variety of human behavior. Another general problem is data sparsity, especially data with high emotional content.

In this paper we evaluate the effects of context change on recognition performance of negative interactions. Our research questions are:

1) what performance can we expect when we deploy a trained system in a new context?
2) what is the most robust modality given context change?
3) can merging data from multiple contexts in the training set mitigate cross-context performance loss?

Our aim is to have a quantification of obtainable performances given context change for the special case of monitoring negative interactions, as would be expected in the afore-mentioned applications. For this purpose, we made audio-visual recordings of interactions in four contexts: at a service desk, a vending machine, in front of lockers and in the cafeteria. We have designed scenarios that are likely to generate negative interactions in each of these contexts. A group of actors was hired for the recordings. They had to interact given a short situation and role description (no scripts), resulting in close to real-life interaction that build up naturally in response to each other's reactions.

Important clues that a situation is escalating can be extracted from the nonverbal behavior of the participants. Aside for very specific behaviors, it is expected that when an interaction becomes problematic people will use wilder body language, more sudden, ample and tense movements. In addition speech has a non-verbal component too, for example changes in pitch, intensity and voice quality. These traits of speech and gestures were named modulation and their relation to stress was explored in [18]. In this paper we use non-verbal (audio-visual) behavior as the indication of negative interactions. To account for context change, we selected audio and video features that are expected to incorporate modulation in the expectation that they are able to generalize better than very specific features like action recognition.

We compared the performance of the selected audio and video features as well as feature level audio-visual fusion in an intra and cross-context scheme. To evaluate whether training exposure to multiple and more diverse situations improves generalization capabilities, we merged data from three contexts for training and tested on the remaining one.

This paper is organized as follows. In section 2 we introduce the collected data in terms of content, procedure, and annotations. We present the experimental setup in section 4, which includes the audio and video features, as well as the classification methodology. In section 5 we present our results, and finalize with conclusions in section 6.

## 2. Data collection

The collected dataset has a direct impact on recognition performance, realism and expected generalization capabilities. One of the much debated issues is the used of acted versus naturalistic datasets [9]. Approaches vary from asking actors to utter predefined texts with different emotions, e.g. [5] to more realistic methods such as Wizard of Oz scenarios of children interacting with a pet robot [1], emotion elicitation by interactions with virtual agents [23], and stress induction by dual-tracking workload computer tasks, or subject motion-fear tasks (subjects in roller-coaster rides) [14]. On the one hand, complaints about using acted data go about the fact that actors tend to exaggerate the emotion portray, and that since the emotion is not real and not spontaneous, the characteristics of the display are different. On the other hand, arguments in favor of using actors given special design considerations acknowledge the fact that real emotions are rare and short lived, and that emotion displays are affected by push (physiologically driven) and pull (social regulation and strategic intention) factors [25]. In the case of recording negative emotions, ethical considerations come into play, which make the recording of real life data challenging. A method that seems to balance the pros and cons of acted and spontaneous recordings is based on improvisations, such as part of the IEMOCAP dataset [6]. For a discussion on advantages and disadvantages of improvised interactions versus scripted interaction please refer to [7].

### 2.1. Content and recording protocol

We have considered four setups (contexts) where problems in interaction are likely to occur: at a service desk, at a vending machine, in a cafeteria and at lockers. For each context, we designed scenarios that are likely to happen and can elicit negative interactions.

A multicultural group of 9 professional actors (4 male, 5 female) were assigned roles based on the considered scenarios. The actors were not given any scripts or specific guidelines besides a short description of the source of conflict and a role. They had to improvise and react to their opponents, which led to close to real-life recordings. Most interactions were between 2 actors, and in few cases there more up to 5 persons in the scene. The spoken languages were Dutch and English, based on the actors' preferences. The scenes were recorded with 2 HD cameras, from two different angles, one of which was used in this study. Each person was wearing a microphone close to mouth, clipped at the shirt.

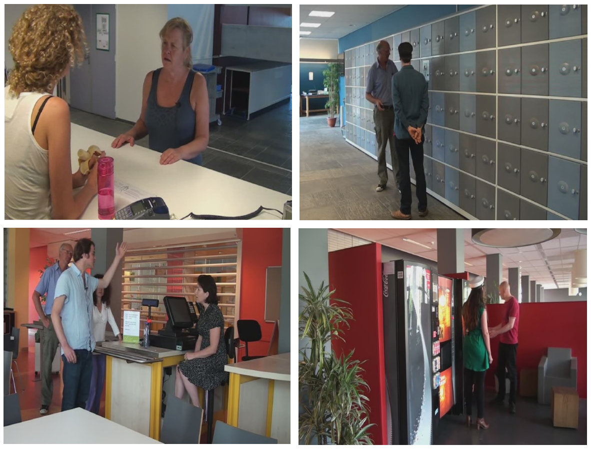

Figure 1. Example footage from the four contexts: service desk, lockers, cafeteria and vending machine.

**2.1.1. Service-desk (SD).** For the service desk (SD) interactions, the actors had to play the roles of service desk employees and customers, given short role descriptions and short instructions. Four scenarios were played two times, resulting in eight sessions, for which the actors did not see the performance of their colleagues from the other session. Example scenarios are a visitor who is late for a meeting and has to deal with a slow employee, a helpless visitor unable to find a location on a map asking the employee to be escorted but being refused, the service desk employee

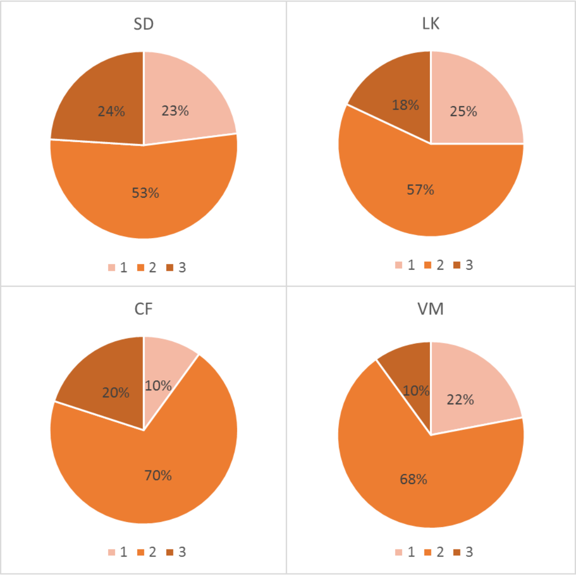

Figure 2. Labels distributions of the data from the four contexts.

does not want to help because of being in lunch break, and the employee or a visitor is in a phone conversation and blocking the service desk.

**2.1.2. Lockers (LK).** The source of conflict in the lockers (LK) scenario is that a person tries to steal from the lockers, pretending the code does not work when a surveillance employee comes. We have recorded two takes of this scenario.

**2.1.3. Vending machine (VM).** In the vending machine (VM) data, and the envisioned scenario is that a costumer plays for an item, but the item does not fall out. When the costumers' dissatisfaction is visible, a staff member comes and there is some interaction between them. We have recorded 4 sessions with this scenario, all of them with different actors.

**2.1.4. Cafeteria (CF).** The last context consists of recordings at the pay desk in a cafeteria (CF). The source of conflict is that costumers want to pay by cash but it is only allowed to pay by card, alternating with a costumer having to load his/her card which takes time and other costumers become impatient. There are four takes in this setting.

Recordings of negative interactions are difficult to obtain in real life: negative events are rare, and ethical and privacy reasons prevent collecting the data. Given these challenges, we believe these datasets are a legitimate choice for our study. The high degree of realism and their similar emotional content increases the suitability and expectations of cross-corpus research.

### 2.2. Annotations

All in all, the selection of contexts illustrated a large variety of situations in which negative interactions occur, or in which at least one of the subjects experiences negative emotions.

The material has been annotated for stress level, on a five point scale, where 1 correspond to a normal situation, and 5 to an extremely negative situation. Note that in this setting, stress is considered from the perspective of surveillance operators. The raters had to evaluate the general stress level in the whole scene, and not per person. The degree of stress was annotated by two to four raters, based on audio-visual data, achieving an agreement of 0.71 measured as Krippendorff's alpha. The annotators were provided with pre-segmented data. The segmentation unit was either utterances, or when not appropriate segments with homogeneous stress level.

For the purpose of the experiments in this paper, we have downsized the annotations to a 3 point scale. More specifically, label 1 was attributed to normal situations, 2 and 3 to moderate stress levels (new label 2), and 4 and 5 to extremely negative situations for which additional help should be provided (new label 3).

## 3. Data processing

### 3.1. Segmentation

Because our aim is to be as close a possible to a real life monitoring situation, we do not assume to know before hand where an utterance begins and ends. Following [20] we decide to analyze segments of equal length, namely two seconds. If a segment spans multiple utterance, it is assigned the label that covers the highest length within those two seconds. Given this setup, we have analyzed a total of 2005 samples, out of which 971 are part of the service desk recordings and represent our training data, and the remaining 267, 340 and 472 representing the lockers, cafeteria and vending machine data. As it would have been the case in real life recordings, the data is unbalanced, with the class representing the most negative situations being the sparsest.

### 3.2. Acoustic features extraction

Given the connection between stress, aggression and negative emotions, we consider acoustic features used in the field of emotion recognition. Popular approaches in speech emotion recognition explore the suprasegmental traits of emotion by applying statistical functionals over the frame level features and using classification/regression techniques on the resulting feature sets [25]. The suprasegmental approaches started off by use of relatively small feature sets, obtained by applying a set of descriptive statistical functionals such as low order moments or extrema to the frame level features [22]. Recently, the brute force approach of feature generation resulting in 1-50k features gained popularity [27]. Feature selection is frequently used to reduce the high dimensionality, but one of the challenges is that the selected features is highly dependent on the chosen corpus [30].

Because of our interest in a small and generic feature set, we chose a feature that had stable performance in a similar cross-corpus study for negative interaction [19]. The software tool Praat [3] was used to extract these features. The feature set consists of the following features: speech duration (without silences), pitch (mean, standard deviation, max, mean slope with and without octave jumps, and range), intensity (mean, standard deviation, max, slope and range), first four formants (F1-F4) (mean and bandwidth), jitter, shimmer, high frequency energy (HF500) (HF1000), harmonics to noise ratio (HNR) (mean and standard deviation), Hammarberg index, spectrum (center of gravity, skewness), and long term averaged spectrum (slope).

### 3.3. Video features extraction

Our work requires low-level video features that are general enough to cover all the manifestations, and discriminative enough to distinguish stress and aggressive events from normal events. We rely on the fact that since we are considering overt manifestation, negative situations are characterized by more gestures, more movements or special characteristics of movement like suddenness [32], [20].

We expect that the most relevant features for stress detection are based on movement. We chose to describe the video segments in terms of space-time interest points (STIP) [16], which are compact representations of the parts of scene which are in motion. Originally these features are employed for action recognition. However they proved suitable for recognizing degrees of aggression [20] and degrees of stress [18]. The space-time interest points are computed for a fixed set of multiple spatio-temporal scales. For the patch corresponding to each interest point, two types of descriptors are computed: histograms of oriented gradient (HOG) to capture appearance, and histograms of optical flow (HOF) to capture movement. These descriptors are used based on a bag-of-words approach, following the approach in [16]. We have computed specialized codebooks, but instead of using K-means as in the original paper, the codebooks were computed in a supervised way using Random Forests with 30 trees and 32 nodes. The resulting feature vectors were reduced using correlation based feature subset selection.

## 4. Classification methodology

In this section we present the experiment setup which includes details on the classification approach, audio and video features, and details on statistical oversampling.

### 4.1. Experiment setup

Our aim is to evaluate cross-context performance and whether it can be improved by training on a combination of multiple contexts, and to check which modalities is more robust to context change. Therefore, while keeping the test dataset fixed, we perform the training on every other dataset and also merge them together for training. For comparison we check the within-corpus performance using 5-fold cross-validation. The experiments are done using audio features, video features and feature level fusion: a concatenation of the audio and video features are used for classification.

Classification is performed using a Random Forest classifier with 100 trees as implemented in Weka [13]. To account for inter-corpus variations, the audio and the video features are normalized per corpus to zero mean and unit standard deviation per feature type. In all cases, given the data unbalance, the evaluation measure is the unweighted average accuracy.

### 4.2. Statistical oversampling

Data unbalance is a frequent problem that affects classification results. There are different possibilities to mitigate this effect, such as adapting the classifier's cost for the under-sampled class, and re-sampling the data to achieve more balance. In this paper we experiment with statistical minority oversampling (SMOTE) [4]. This method generates artificial new samples of the minority class by adding noise to the data. The percentage of new generated data is a parameter that has to be set. We apply SMOTE only on the training set. Based on the initial label distribution of the train data, we have applied statistical oversampling for the two least represented classes, with a precomputed parameter to even out the distributions.

## 5. Results and discussion

We present results in terms of average unweighted accuracies to account for original unbalance in the data. Figures 3-6 present results on testing on each of the four contexts respectively. The results plotted on the left part of each figure are obtained with training on the same context (within corpus training and testing with 5-fold cross-validation). They are followed by the three cross-context results, and on the right part each figure displays the results for training on a concatenation of all three contexts. The three lines in each figure correspond to using audio, video and feature-level fusion.

The results clearly indicate that the best performance is achieved for the within context setup. A performance drop is observed for all cross-context cases. Intuitively, one might expect that training a system on a combination of contexts, instead of on only one context, increases robustness. However, our results do not support this expectation. Interestingly, best cross-corpus results tend to be obtained when training was done on the VM context.

Another consistent finding for all test-contexts is that the audio features outperformed the video ones. Audio-visual fusion yields accuracies very close to the ones of audio, but rarely outperforms audio for the cross-context cases. Interestingly, in the case of merging the three (other) contexts for training, audio-visual fusion does improve over audio.

The lower performance of video compared to audio can be influenced by the fact that visible human behavior spans a much higher range in between and even within the same context, compared to audible behavior. For example, the service desk dataset is characterized by a close-up view to the actors, and a lot of hand gestures. Because some of the scenarios are about time pressure, the recordings contain gestures such as checking time, fidgeting, but also conversational gestures amplified by negative affect. The other three test sets contain recordings from a higher distance. In the cafeteria data there are more people visible, and there is more movement such as walking, since frequently the actors went to charge their card. The vending machine data contains more violent movements, e.g. tendencies to vandalize the vending machine. In the lockers data there was less restriction about where the actions is going to take place, and there is more occlusion. There are changes in audio characteristics as well, such as different room acoustics and noise levels. The recordings were made during normal working hours and normal conditions, therefore containing the recording quality that can be expected in deploying a system in a real situation.

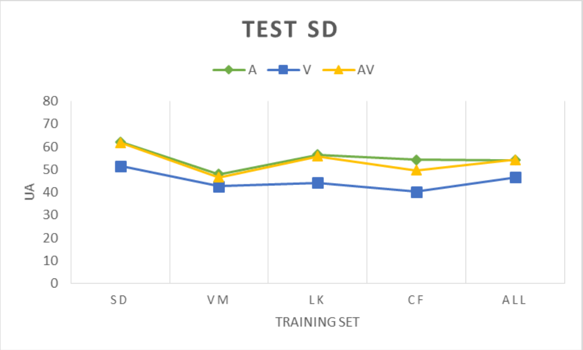

Figure 3. Results in unweighted accuracies for testing on the Service Desk context, when training is done from left to right on Service Desk (SD), Vending Machine (VM), Lockers (LK), Cafeteria (CF), and on a merged set of VM+LK+CF (all). The three lines correspond to results using audio feature (A), video features (V) and audio-visual feature level fusion (AV).

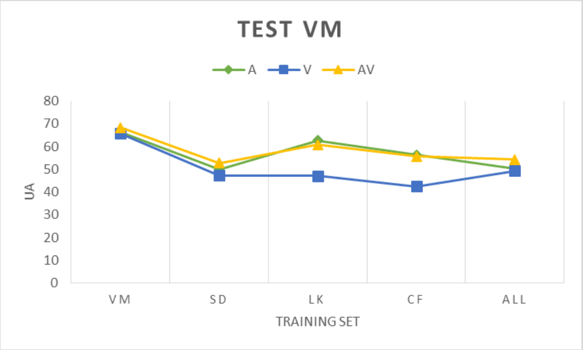

Figure 4. Results in unweighted accuracies for testing on the Vending Machine (VM) context, when training is done from left to right on Vending Machine (VM), Service Desk (SD), Lockers (LK), Cafeteria (CF), and on a merged set of SD+LK+CF (all). The three lines correspond to results using audio feature (A), video features (V) and audio-visual feature level fusion (AV).

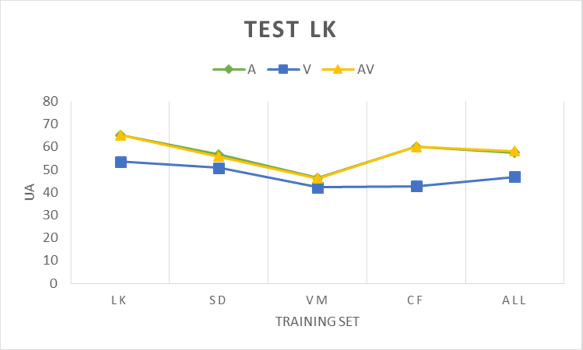

Figure 5. Results in unweighted accuracies for testing on the Lockers (LK) context, when training is done from left to right on Lockers (LK), Service Desk (SD), Vending Machine (VM), Cafeteria (CF), and on a merged set of LK+SD+CF (all). The three lines correspond to results using audio feature (A), video features (V) and audio-visual feature level fusion (AV).

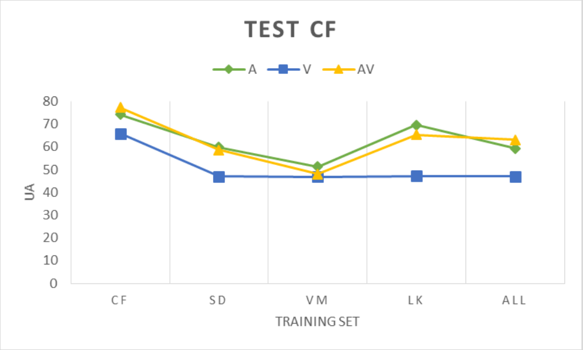

Figure 6. Results in unweighted accuracies for testing on the Cafeteria (CF) context, when training is done from left to right on Cafeteria (CF), Service Desk (SD), Vending Machine (VM), Lockers (LK),and on a merged set of SD+VM+LK (all). The three lines correspond to results using audio feature (A), video features (V) and audio-visual feature level fusion (AV).

## 6. Conclusion

One of the challenges of deploying trained recognition systems in real-life is that they are always trained with a limited set of situations and they should be able to generalize well to new, unseen situations. The considered application is surveillance of human-human interaction in situations that might escalate. It is important to recognize timely the negative interaction and offer support such that the situation will not get out of hand. In this paper we focused on evaluating the performance of detecting negative interactions when the test set and the training set are from different contexts. Our approach was to focus on nonverbal behavior and on behavior traits that are expected to be similar disregarding the considered context such as speech modulation (changes in pitch, intensity, voice quality) and body language modulation (speed, rhythm, repetition).

Coming back to answering the first research questions formulated in the introduction, our results consistently show that when testing is done on a context different form the training one, there is a performance drop compared to intra-corpus training and testing. This finding was to be expected given previous studies from the speech community. Differences in amount of decrease can perhaps be attributed to context (dis)similarities. Interestingly, there was one context (Lockers) that proved to be best for training when the other contexts are used for testing.

The second research question focused on the best performing modality in context change. Our results show that speech was consistently the best predictive modality, followed closely by audio-visual fusion. Intuitively, there are less differences in how people speak during negative interactions in different contexts, given that we consider only the paralinguistic part of speech (not the semantics). On the contrary, there are significant differences in the video content for the four considered contexts, ranging from camera angle view and distance to the participants to the expected behaviors, which can be an explanation of our findings.

Finally, to answer the last research question we experimented training on merged sets of three contexts and testing on the remaining one. The idea was that giving the system more examples of situations and behaviors, it will be able to generalize better. However the results show that training on a merged set achieves a performance slightly lower than the best performing single cross-context. We can probably attribute this finding to the fact that some contexts are more similar to the test context, resulting in better performance.

## References

[1] Batliner, A., Steidl, S., Nöth, E.: Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus. In: Proc. of a Satellite Workshop of LREC, pp. 28–31 (2008)

[2] Berkowitz, L.: Affective aggression: The role of stress, pain, and negative affect. (1998)

[3] Boersma, P.: Praat, a system for doing phonetics by computer. Glot International **5**(9/10) (2001)

[4] Bowyer, K.W., Chawla, N.V., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. CoRR **abs/1106.1813** (2011)

[5] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of german emotional speech. In: Interspeech, vol. 5, pp. 1517–1520 (2005)

[6] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., Narayanan, S.: Iemocap: interactive emotional dyadic motion capture database. Language Resources and Evaluation **42**(4), 335–359 (2008)

[7] Busso, C., Narayanan, S.S.: Scripted dialogs versus improvisation: lessons learned about emotional elicitation techniques from the iemocap database. In: Interspeech, pp. 1670–1673 (2008)

[8] Cao, L., Liu, Z., Huang, T.: Cross-dataset action detection. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 1998–2005 (2010)

[9] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Towards a new generation of databases. Speech communication **40**(1), 33–60 (2003)

[10] Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P.: Design, recording and verification of a danish emotional speech database. In: Eurospeech (1997)

[11] Esposito, A., Esposito, A.M., Martone, R., Müller, V., Scarpetta, G.: Towards Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues: Third COST 2102 International Training School, Caserta, Italy, March 15-19, 2010, Revised Selected Papers, vol. 6456. Springer Science & Business Media (2011)

[12] Grimm, M., Kroschel, K., Narayanan, S.: The vera am mittag german audio-visual emotional speech database. In: Multimedia and Expo, 2008 IEEE International Conference on, pp. 865–868. IEEE (2008)

[13] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter **11**(1), 10–18 (2009)

[14] Hansen, J., Bou-Ghazale, S., Sarikaya, R., Pellom, B.: Getting started with susas: a speech under simulated and actual stress database. In: EUROSPEECH, vol. 97, pp. 1743–46 (1997)

[15] Kim, S., Valente, F., Vinciarelli, A.: Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 5089–5092 (2012)

[16] Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Int. Conf. of Computer Vision and Pattern Recognition (2008)

[17] Lefter, I., Burghouts, G., Rothkrantz, L.: An audio-visual dataset of human-human interactions in stressful situations. Journal on Multimodal User Interfaces **8**(1), 29–41 (2014)

[18] Lefter, I., Burghouts, G.J., Rothkrantz, L.J.M.: Recognizing stress using semantics and modulation of speech and gestures. IEEE Transactions on Affective Computing **7**(2), 162–175 (2016)

[19] Lefter, I., Nefs, H.T., Jonker, C.M., Rothkrantz, L.J.M.: Cross-corpus analysis for acoustic recognition of negative interactions. In: International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 132–138 (2015)

[20] Lefter, I., Rothkrantz, L., Burghouts, G.: A comparative study on automatic audio-visual fusion for aggression detection using meta-information. Pattern Recognition Letters **34**(15), 1953 – 1963 (2013)

[21] Lefter, I., Rothkrantz, L.J., Van Leeuwen, D.A., Wiggers, P.: Automatic stress detection in emergency (telephone) calls. International Journal of Intelligent Defence Support Systems **4**(2), 148–168 (2011)

[22] Li, Y., Zhao, Y.: Recognizing emotions in speech using short-term and long-term features. In: ICSLP (1998)

[23] McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. Affective Computing, IEEE Transactions on **3**(1), 5–17 (2012)

[24] Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A., et al.: Challenges of human behavior understanding. In: HBU, pp. 1–12. Springer (2010)

[25] Scherer, K.R., Bänziger, T.: On the use of actor portrayals in research on emotional expression. In: K.R. Scherer, T. Bänziger, E.B. Roesch (eds.) Blueprint for affective computing: A sourcebook,, pp. 166–176. Oxford, England: Oxford university Press (2010)

[26] Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.: Cross-corpus acoustic emotion recognition: Variances and strategies. Affective Computing, IEEE Transactions on **1**(2), 119–131 (2010)

[27] Schuller, B., Wimmer, M., Mosenlechner, L., Kern, C., Arsic, D., Rigoll, G.: Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? In: Acoustics, Speech and Signal Processing (ICASSP) IEEE International Conference on, pp. 4501–4504. IEEE (2008)

[28] Sprague, J., Verona, E., Kalkhoff, W., Kilmer, A.: Moderators and mediators of the stress-aggression relationship: Executive function and state anger. Emotion **11**(1), 61–73 (2011)

[29] Tahon, M., Delaborde, A., Devillers, L.: In: INTERSPEECH, pp. 3121–3124. ISCA

[30] Vogt, T., Andre, E.: Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Multimedia and Expo (ICME), IEEE International Conference on, pp. 474–477 (2005)

[31] Weninger, F., Schuller, B.: Discrimination of linguistic and non-linguistic vocalizations in spontaneous speech: Intra-and inter-corpus perspectives. In: INTERSPEECH (2012)

[32] Zajdel, W., Krijnders, J.D., Andringa, T., Gavrila, D.M.: Cassandra: audio-video sensor fusion for aggression detection. In: Advanced Video and Signal Based Surveillance, AVSS. IEEE Conference on, pp. 200–205. IEEE (2007)