Detection of Challenging Dialogue Stages Using Acoustic Signals and Biosignals

Olga Egorow
IIKT
Otto von Guericke University
39106 Magdeburg
Germany
olga.egorow@ovgu.de

Andreas Wendemuth
IIKT & CBBS
Otto von Guericke University
39106 Magdeburg
Germany
andreas.wendemuth@ovgu.de

ABSTRACT

Emotions play an important role in human-human interaction. But they are also expressed during human-computer interaction, and thus should be recognised and responded to in an appropriate way. Therefore, emotion recognition is an important feature that should be integrated in human-computer interaction. But the task of emotion recognition is not an easy one – in "in the wild" scenarios, the occurring emotions are rarely expressive and clear. Different emotions like joy and surprise often occur simultaneously or in a very reduced form. That is why, besides recognising categorial and clear emotions like joy and anger, it is also important to recognise more subtle affects. One example for such an affect that is crucial for human-computer interaction is trouble experienced by the human in case of unexpected dialogue course. Another point concerning this task is that the emotional status of a person is not necessarily revealed in his or her voice. But the same information is contained in the physiological reactions of the person, that are much harder to conceal, therefore representing the "true signal". That is why the physiological signals, or biosignals, should not be left unattended. In this paper we use the data from naturalistic human-computer dialogues containing challenging dialogue stages to show that it is possible to differentiate between troubled and untroubled dialogue in acoustic as well as in physiological signals. We achieve an unweighted average recall (UAR) of 64% using the acoustic signal, and an UAR of 88% using the biosignals.

Keywords

Emotion, affect, affective computing, emotion recognition, acoustic emotion recognition, biosignals

1 INTRODUCTION

One of the goals of human-computer spoken interaction is to become more and more similar to human-human interaction, turning the machine into an almost-human companion. For this matter, not only understanding what a human is saying is important, but also how it is said – the emotions of the human counterpart are an equally important part of interaction. This is why computers should be able to recognise and understand emotions. But this is a challenging task, since human emotions can range in a variety of dimensions. As a starting point, we can consider six basic emotions following the categorial model [1]: happiness, surprise, fear, sadness, anger and disgust combined with contempt. But natural emotions comprise clearly more than that. Another classification of emotions, better suited for natural emotions, is the multidimensional scale of pleasure-arousal-dominance [2]. But not all emotions are relevant for human-computer interaction. One emotion is especially important in this aspect: the feeling that humans experience when something unexpected happens during a dialogue. If an interaction suddenly becomes challenging for the human counterpart, it is crucial for the success of the dialogue to recognise it and to react in an appropriate way.

In this paper we show that even slight emotional changes occurring in naturalistic human-machine dialogue can be predicted from acoustic as well as from physiological signals. For this purpose, we use data of human-computer interaction obtained during a naturalistic multimodal Wizard-of-Oz (WOZ) experiment, consisting of unchallenging and challenging dialogue stages. This design allows us to look into problems that naturally arise from challenging human-machine interaction. For this purpose, we simplify the dialogue stages to two main classes: the *baseline* class containing normal interaction and the *trouble* class containing challenging interaction. We show that it is possible to automatically

discriminate these dialogue stages based on acoustic data as well as on biosignals.

The remainder of the paper is structured in the following manner: first, we give an overview on existing related work in section 2; in section 3 we introduce the data set used for our experiments; in section 4 we describe the feature extraction routines for both data types, and the classification process; in section 5 we present and discuss the achieved results; in section 6 we summarise the work and show some possibilities for future development.

2 RELATED WORK

Automatic emotion recognition on acoustic data has achieved considerable results in the last years. On acted data, like the standard Berlin Database of Emotional Speech [3], recognition rates of 85% are possible [4]. But in realistic tasks, emotions often cannot be divided in discrete categories of joy, sadness, anger, etc. The automatic classification of naturalistic data is difficult and achieves less impressive outcomes: for "in the wild" scenarios using audio only recognition results of about 33% UAR [5] to about 24% UAR [4] and 20% accuracy [6] for seven-class-problems are achieved. But even for human annotators emotion recognition from speech alone is not an easy task – for acoustic data without context, an average accuracy of 60% is achieved by human raters [7]. To solve this problem, humans rely on other modalities of interaction besides speech, like facial expressions, gestures, etc. This applies also to automatic emotion recognition. For example, adding the dimension of co-speech gestures improves the results by 2 percentage points compared to speech alone [8]. Another experiment shows that using only facial expressions outperforms using only acoustics by almost 15 percentage points in terms of accuracy, and using both data types simultaneously in a bimodal system leads to a further improvement of 4 percentage points [9].

Additionally, there are physiological reactions related to emotions, like a racing heart, gasping and sweating palms. The links between biosignals such as heart rate, temperature, skin resistance and emotions have been known for a long time [10] [11] [12]. One big advantage of using physiological data is that biosignals can be obtained at all times – acoustic signals can only be obtained when the user is speaking. Another point is that physiological reactions, in contrast to reactions in voice, are much harder to conceal. That is why it is not surprising, that recognition of emotions on physiological signals achieves better results than only on acoustic data. This has been proven many times, for example for categorial emotions, where the recognition rate ranges from 49% to 75% for induced emotions [13] [14] [15], for the multidimensional valence-arousal scale with a recognition rate of over 90% [16], also for induced emotions, and for other scales like fun levels, with a recognition rate of 70% [17].

There are also multimodal approaches combining both physiological and acoustic data. One example is similar to our setting: a WOZ quiz show with four stages, corresponding to four classes on the valence-arousal scale (high and low for valence and arousal, respectively), where six different biosignals and acoustic feedback are used for emotion recognition [18]. The results vary depending on the evaluation method (92%-69% accuracy for subject-dependent vs. 55% accuracy for subject-independent evaluation), but combining the physiological and acoustic features leads to an improvement in all cases. Although this direction seems promising, the research on this topic is still rare. One possible problem is that the gathering of naturalistic multimodal data is not a simple process, in terms of recording (e.g. problems concerning the synchronisation of data streams) as well as in terms of processing (e.g. fusion of data streams).

Most of the studies presented above deal with emotion recognition in general and investigate elicited emotions. But on the important topic of recognising trouble in human-machine communication, especially naturalistic communication, not much research has been done so far. The groundbreaking example is detecting trouble in acted and elicited interaction using acoustic data and detailed annotations containing part-of-speech (POS) tagging, dialogue acts, repetitions and syntactic boundaries, achieving 73% to 96% recall for different scenarios [19]. One of the scenarios described in this approach is a naturalistic WOZ experiment, here the data was separated into two classes: one class containing prosodic peculiarities and one class containing no prosodic peculiarities. This setup resembles the setup of our study, but in our case we rely on much simpler annotations of the acoustic data and, more importantly, on biosignals.

3 THE DATA SET

3.1 The LAST MINUTE Corpus

The LAST MINUTE Corpus [20][21][22] contains naturalistic multimodal recordings of German speaking subjects in a WOZ experiment. The setup of the experiment revolves around the preparations for an imaginary journey to an unknown place "Waiuku". Each experiment lasts about 30 minutes and consists of several dialogue stages, each triggered by a major event. A summary of the different dialogue stages can be seen in Table 1. First, the subjects are asked to introduce themselves to the "machine" in a "warm-up" dialogue stage. After that, they are requested to imagine winning a summer trip to an unknown destination called Waiuku, and they have to pack a suitcase by choosing items from a list in a "listing" dialogue stage. There is also a time constraint: the trip begins immediately,

and the subjects have only fifteen minutes for the packing process. After several minutes of packing there is another major event: the subjects learn that the suitcase has a weight limit, so they have to remove some of the packed items. This corresponds to the "challenge" dialogue stage. After that, the next major event occurs when the real destination of the trip is revealed: Waiuku lies in the southern hemisphere. Since the subjects now know that the trip is a winter trip instead of a summer trip, they have to re-organise their suitcase again. This dialogue stage is called the "Waiuku" stage. At the end of the experiment, there is a short "conclusion" stage. It is expected that the subjects experience different emotions during the dialogue stages, and also express them. It should be noted that like in any naturalistic scenario, the subjects may react differently, with reactions ranging from very expressive to very subtle.

Dialogue Stage	Trigger	Troubled?
Warm-up	Introduction request	No
Listing	Winning the trip	No
Challenge	Weight Constraint	Yes
Waiuku	Revealing destination	Yes
Conclusion	Concluding remarks	No

Table 1: Overview of the dialogue stages

The acoustic data is recorded using 2 directional microphones at 44100 Hz and stored in the wav format. The biosignal data is recorded using the NeXus-32 system[1]. From the various biosignals available from this system, it proved sufficient for our analysis to use electromyogram (EMG), skin conductivity (SC) and respiration (RSP). These biosignals could be obtained in sustained quality throughout the experiment.

3.2 Selecting the Data

From all the recordings of the LAST MINUTE Corpus we selected a subset containing the recordings of 19 subjects, of whom both the acoustic and the physiological data exist. The age and sex distribution of the subjects is nearly balanced, as shown in Table 2. The acoustic data set and the biosignal data set are divided into three subsets to enable subject-independent evaluation. The subsets for the acoustic data contain the same subjects as the subsets of the biosignal data, leading to a training subset containing data of 11 subjects, a development subset containing data of 4 subjects and a test subset containing data of 4 subjects each. These subsets are also nearly balanced regarding the distribution of age and sex, cf. Table 2. In the classification process, the models are built on the training subsets, fine-tuning the parameters of the classifier takes place on the development subsets in order to avoid overfitting to the test data, the test subsets are used to obtain the final classification results.

	Train	Dev	Test	Overall
Sex				
Female	5	2	2	9
Male	6	2	2	10
Age				
Young (< 30)	7	2	2	11
Elder (> 60)	4	2	2	8

Table 2: Distribution of sex and age of the subjects.

3.3 Dividing the Data into Classes

The hypothesis of this paper is that it is possible to automatically detect the different dialogue stages described above in both acoustic and biosignal data recorded during the experiments. For this purpose, we divide the data into two classes: the *baseline* class, denoting untroubled interaction in the warm-up, listing and conclusion dialogue stages, and the *trouble* class, denoting the challenge and Waiuku dialogue stages, where the subjects are expected to experience trouble during the interaction. It should be noted that no perception tests were conducted, therefore the data is not annotated regarding the level of trouble expressed by the subject. The labels consist only of the dialogue stages. Therefore, the trouble class contains different levels of trouble. We will present the different levels using two examples from the challenge dialogue stage.

The first example shows two snippets from a dialogue, here the subject is an elderly woman. In the first part, she is – for the first time – informed by the Wizard that the weight limit is reached:

Wizard: *A swimsuit or bikini cannot be added, otherwise the maximum weight limit prescribed by the airline would be exceeded. Before other items can be selected, you must provide enough space in your suitcase. For this, already packed items can be unpacked. On demand, you can get a list of the already selected items.*

Subject: *Yes, uh ((pause)) I would like ((pause)) take out a pair of shoes.*

Wizard: *Your statement cannot be processed.*

After she fails to remove some items, she selects some more items and gets the same message from the Wizard again. Now she seems frustrated:

Wizard: *Before other items can be selected, you must provide enough space in your suitcase. For this, already packed items can be unpacked. On demand, you can get a list of the already selected items.*

[1] http://www.mindmedia.info/CMS2014/products/systems/nexus-32

Subject: *hm ((moaning)) yes (.) then I want to hear the chosen items again please, I told you I want to unpack shoes.*

The second example shows the dialogue of a young woman, who also learns that there is a weight limit:

Wizard: *Before other items can be selected, you must provide enough space in your suitcase. For this, already packed items can be unpacked. On demand, you can get a list of the already selected items.*

Subject: *Remove inflatable boat.*

Wizard: *One inflatable boat was removed, you can continue.*

Subject: *((smacks)) three bikinis ((swallows))*

She seems to be less influenced by the weight limit, at least concerning her speech alone.

Both dialogue snippets are examples of the *trouble* class, since both parts happen during the challenge dialogue stage.

4 RECOGNITION EXPERIMENTS

4.1 Pre-processing the Data

The acoustic feature set consists of the Emobase feature set, fully described in [23], which is widely used for emotion recognition. The features are extracted using openSMILE [24]. The feature set includes 988 acoustic features extracted on utterance-level, such as intensity, loudness, 12 MFCCs, F_0, voicing probability F_0 envelope, 8 line spectral frequencies, zero-crossing rate, and their functionals. Other feature sets widely employed for emotion recognition, such as those described in [25] [26] were tested in a preliminary investigation, but were rejected since they lead to poor results.

The physiological features are extracted from the biosignal data on dialogue stage level (including the speaking time of the Wizard), using the Augsburg Biosignal Toolbox[2]. The 3 original biosignals (EMG, RSP, SC) are preprocessed by applying a lowpass filter and normalisation, then a total number of 104 features, including first and second order derivatives and statistical features (mean, median, standard deviation, etc.) are calculated at a sampling rate of 32 Hz. The full description of the feature set can be found in [27].

4.2 Classification

We chose random forest as a classifier for the classification of both, acoustic and biosignal data. This classification method was chosen because of its higher training speed and its good performance compared to support vector machines [28], the standard classification method widely used for emotion recognition from speech. We employ the Weka implementation of random forest, which is based upon the classic algorithm by Breiman [29]. One advantage of this implementation is that there are only two parameters to be tuned: the number of features used in each node and the number of trees. The hyperparameter optimisation takes place using grid search. For both types of data, between 1 and 50 features and between 10 and 100 trees are evaluated using the development subsets. For the acoustic data, the best parameters are found to be 6 features and 30 trees. For the biosignal data, the best parameters are found to be 3 features and 10 trees.

[2] `http://www.informatik.uni-augsburg.de/lehrstuehle/hcm/projects/tools/aubt/`

5 RESULTS AND DISCUSSION

The recall, precision and f-measure of the classification for the two classes of *trouble* and *baseline* are shown in Table 3 and Table 4 for the acoustic and biosignal data, respectively. A comparison of the UAR values for both types of data can be seen in Fig. 1.

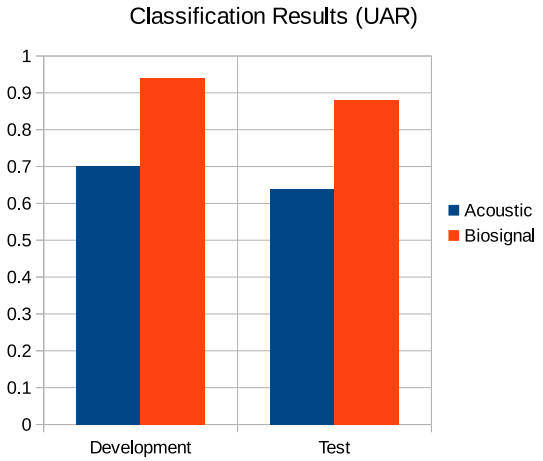

Figure 1: Comparison of classification results on physiological and acoustic data, UAR

On acoustic data, the UAR lies at 0.70 for the development set and 0.64 for the test set, with higher recall values for the *trouble* class and lower values for the *baseline* class. On biosignal data, the results are better by roughly 25 percentage points: the UAR lies at 0.94 for the development set and 0.88 for the test set, here the *trouble* class is recognised with a higher recall compared to the *baseline* class on the development set, but with a lower recall on the test set. But overall we can see that the results on the test set are similar to the results on the development set, indicating that the model is able to appropriately generalise.

Regarding the precision of the detection we can see a comparable trend as for UAR. For the acoustic data, the unweighted average precision lies at 0.68 and 0.64 for the development and the test sets, respectively. For the

biosignal data, the values of unweighted average precision are 22 percentage points higher for the development set and even 30 percentage points higher for the test set, resulting in 0.90 for the development set and an even higher value of 0.94 for the test set.

	Recall	Precision	F-Measure
Development Set			
Trouble	0.77	0.51	0.61
Baseline	0.63	0.84	0.72
Unweighted av.	0.70	0.68	0.67
Test Set			
Trouble	0.66	0.58	0.62
Baseline	0.62	0.70	0.66
Unweighted av.	0.64	0.64	0.64

Table 3: Classification results on acoustic data.

	Recall	Precision	F-Measure
Development Set			
Trouble	1.00	0.80	0.89
Baseline	0.88	1.00	0.93
Unweighted av.	0.94	0.90	0.91
Test Set			
Trouble	0.75	1.00	0.86
Baseline	1.00	0.89	0.94
Unweighted av.	0.88	0.94	0.90

Table 4: Classification results on physiological data.

Overall we can say that *trouble* can be recognised in both, the biosignal and the acoustic data, but the classification on the biosignal data clearly outperforms the classification on the acoustic data. This can be explained by the fact that, as already mentioned, the emotions contained in voice are easy to conceal, in contrast to the physiological reactions, which cannot be controlled deliberately.

Comparing our results to those found in the literature, we can say that the results on acoustic data are not as good as presented in [19], where an average recall of over 73% for a WOZ scenario and a two class problem (prosodic peculiarities vs. no prosodic peculiarities during a challenging dialogue) was reached. But as already mentioned, the data used there had a more detailed annotation, including POS tagging and, more importantly, annotations of prosodic peculiarities detected by the annotators, and not only annotations of dialogue stages supposed to lead to trouble, as in our case. Concerning biosignals, we achieve better results than the results described in [18]. In a comparable WOZ setting including four levels on the valence-arousal scale, only 55% accuracy in subject-independent evaluation combining acoustic and biosignal data can be achieved there.

In general, we can say that recognising challenging stages of dialogues using biosignal data is reliable: even in this subject-independent evaluation we can recognise the *trouble* class with a very high certainty - we found 75% of all instances of the *trouble* class, with 0% false alarm rate. Unfortunately, the same cannot be said for using the acoustic data. For this case, we found only 58% of the instances, and only 70% of the found instances were indeed instances of the *trouble* class.

Although the results for the biosignal data are very promising, we have to consider that this data, in contrast to acoustic data, is not easily obtainable, especially EMG and RSP. We can assume that the compliance of human-computer interaction systems might suffer from intrusiveness of physiological sensors (here intrusiveness means constraints to the observed human). On the other hand, there are also easily obtainable types of physiological data, such as pulse and skin temperature, which can be collected from interaction devices like smartwatches etc. For further investigations, it would be interesting to focus on these easily obtainable types of physiological data.

One probable explanation for the different results on acoustic and biosignal data is that, as already mentioned, acoustic data can be easily manipulated by the subject. It is imaginable that some of the subjects forced themselves to speak calmly, since they were speaking to a computer. But in contrast to voice, the physiological reactions cannot be deliberately manipulated, and therefore more differences between the dialogue stages can be found and thus automatically detected. This also means, that the biosignal data can be used as "ground truth" to detect changes in human emotional state that cannot be detected from speech, and also to annotate them. But, on the other hand, trouble recognition using only acoustics also should not be ignored: for tasks where no data other than acoustic data is available we can still detect over 60% of challenging dialogue stages using our approach. One of such tasks could be call center applications, where a trouble detection system could support human call center agents [30].

Another problem concerning emotion recognition from speech is that there is still no consensus in the literature regarding which features are best suited for this difficult task – it might be that the usually employed features do not represent the differences between various emotions. Additionally, many feature extraction routines base on human perception models – but, as already mentioned before, emotion recognition cannot be done with a 100% accuracy by human annotators. This also opens the question of gathering the "golden stan-

dard" – to build the right model for emotion recognition, we need to ensure that the emotions are labelled correctly in the data. A widely employed but costly solution for this problem is to obtain annotations from multiple raters and to use only data with a high interrater agreement, which, however, is also difficult to achieve [31]. Our results encourage to rely on biosignal data as ground truth instead, therefore saving the effort of multiple annotation procedures.

6 CONCLUSION

In this paper, we investigated how challenging dialogue stages in naturalistic human-computer interaction can be automatically recognised. For this task, we used the recordings of the LAST MINUTE Corpus. The recordings include non-challenging and challenging parts of WOZ human-computer interaction, which were consolidated into two classes: the *baseline* class and the *trouble* class. Instead of widely employed support vector machines we used random forest for this classification task. We achieved an UAR of 64% on acoustic data and an UAR of 88% on biosignal data, showing that it is possible to detect challenging parts of an interaction using acoustic data as well as physiological data. However, we did not perform human perception tests regarding the levels of trouble audible in the acoustic data and used only simple annotations.

There are two main directions for future research. First, it should be investigated, whether the recognition rate can be improved by more complex annotations of different levels of trouble. As mentioned before, different subjects may experience and express different levels of trouble. An important question is whether age, sex or other factors influence the experienced and expressed level of trouble during a challenging human-computer interaction. If this is the case, using different models for different user groups should improve the results.

Another direction for future work is to exploit the multimodality of the data, using both acoustic and biosignal data simultaneously, since it was already proven in the literature that multimodal approaches can improve the detection results [32]. Especially combinations of acoustics and easily obtainable biosignals like pulse could be interesting for this task. Unfortunately, a multimodal investigation was not possible in this setting because of missing synchronisation of the used data sets. We will approach this problem in future research.

7 ACKNOWLEDGEMENT

The work presented in this paper was done within the Transregional Collaborative Research Centre SFB/TRR 62 'Companion-Technology for Cognitive Technical Systems' (`www.sfb-trr-62.de`) funded by the German Research Foundation (DFG). The first author was additionally funded by the consortium 3Dsensation (`www.3d-sensation.de/`), a part of the Zwanzig20 German government funding program.

8 REFERENCES

[1] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings.* New York: Pergamon Press, 1972.

[2] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, pp. 261–292, 1996.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of the INTERSPEECH-2005*, Lisbon, Portugal, 2005, pp. 1517–1520.

[4] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proceedings of the IEEE ASRU-2009*, Merano, Italy, 2009, pp. 552–557.

[5] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller, "Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion," in *Proceedings of the 16th ICMI*. ACM, 2014, pp. 473–480.

[6] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proceedings of the 15th ICMI*. ACM, 2013, pp. 517–524.

[7] K. R. Scherer, "Speech and emotional states," *Speech evaluation in psychiatry*, pp. 189–220, 1981.

[8] R. Böck, K. Bergmann, and P. Jaecks, "Disposition recognition from spontaneous speech towards a combination with co-speech gestures," in *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*. Springer, 2014, pp. 57–66.

[9] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th ICMI*. ACM, 2004, pp. 205–211.

[10] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[11] R. W. Levenson, P. Ekman, and W. V. Friesen, "Voluntary facial action generates emotion-specific autonomic nervous system activity," *Psychophysiology*, vol. 27, no. 4, pp. 363–384, 1990.

[12] J. T. Cacioppo, G. G. Berntson, J. T. Larsen, K. M. Poehlmann, and T. A. Ito, "The psychophysiology of emotion," *Handbook of emotions*, vol. 2, pp. 173–191, 2000.

[13] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, "Basic emotions are associated with distinct patterns of cardiorespiratory activity," *International journal of psychophysiology*, vol. 61, no. 1, pp. 5–18, 2006.

[14] K. H. Kim, S. Bang, and S. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.

[15] A. Choi and W. Woo, "Physiological sensing and feature extraction for emotion recognition by exploiting acupuncture spots," in *Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 590–597.

[16] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *Affective Dialogue Systems*. Springer, 2004, pp. 36–48.

[17] G. N. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," *International Journal of Human-Computer Studies*, vol. 66, no. 10, pp. 741–755, 2008.

[18] J. Kim, "Bimodal emotion recognition using speech and physiological changes," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. Vienna, Austria: I-Tech Education and Publishing, 2007, pp. 265–280.

[19] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Communication*, vol. 40, pp. 117–143, 2003.

[20] J. Frommer, B. Michaelis, D. Rösner, A. Wendemuth, R. Friesen, M. Haase, M. Kunze, R. Andrich, J. Lange, A. Panning, and I. Siegert, "Towards emotion and affect detection in the multimodal last minute corpus," in *Proceedings of the 8th LREC*, Istanbul, Turkey, 2012, pp. 3064–3069.

[21] D. Rösner, J. Frommer, R. Friesen, M. Haase, J. Lange, and M. Otto, "LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions," in *Proceedings of the 8th LREC*, Istanbul, Turkey, 2012, pp. 96–103.

[22] D. Prylipko, D. Rösner, I. Siegert, S. Günther, R. Friesen, M. Haase, B. Vlasenko, and A. Wendemuth, "Analysis of significant dialog events in realistic human-computer interaction," *Journal on Multimodal User Interfaces*, vol. 8, pp. 75–86, 2014.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. of the ACM MM-2010*, Firenze, Italy, 2010, pp. 1459–1462.

[24] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838.

[25] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proceedings of the INTERSPEECH-2009*, Brighton, UK, 2009, pp. 312–315.

[26] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proceedings of the INTERSPEECH-2011*, Florence, Italy, 2011, pp. 3201–3204.

[27] J. Wagner, *The Augsburg biosignal toolbox*. University of Augsburg, 2009.

[28] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[29] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] I. Siegert and K. Ohnemus, "A new dataset of telephone-based human-human call-center interaction with emotional evaluation," in *Proceedings of the 1st International Symposium on Companion Technology*, Ulm, Germany, September 2015, pp. 143–148.

[31] I. Siegert, R. Böck, and A. Wendemuth, "Inter-rater reliability for emotion annotation in human-computer interaction: comparison and methodological improvements," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 17–28, 2014.

[32] F. Schwenker, S. Scherer, and L. Morency, "Multimodal pattern recognition of social signals in human-computer-interaction," *Lecture Notes in Computer Science*, vol. 8869, 2015.