Egorow, Olga; Siegert, Ingo; Wendemuth, Andreas:

Prediction of User Satisfaction in Naturalistic Human-Computer Interaction

In: Kognitive Systeme / 2017 - 1

DOI: http://dx.doi.org/10.17185/duepublico/44534

URN: urn:nbn:de:hbz:464-20170926-155216-5

Link: http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=44534

Prediction of User Satisfaction in Naturalistic Human-Computer Interaction

Olga Egorow*. Ingo Siegert*. Andreas Wendemuth*

** Cognitive Systems Group, Otto-von-Guericke University Magdeburg,*
39016 Magdeburg, Germany
(e-mail: <firstname>.<lastname>@ovgu.de)

Abstract: User satisfaction is an important aspect of human-computer interaction (HCI) – if a user is not satisfied, he or she might not be willing to use such a system. Therefore, it is crucial to HCI applications to be able to recognise the user satisfaction level in order to react in an appropriate way. For such recognition tasks, data-driven methods have proven to deliver useful and robust results. But a data-driven user satisfaction model needs labelled and reliable data, which is not always easy in the case of user satisfaction, since it is not accessible directly. In the investigation presented here, the users are asked directly about their satisfaction level regarding their performance in a task during a close-to-real-life HCI. This results in a one-to-one mapping between a satisfaction level and the expressed vocal characteristics in the user's utterance. This data is then used to build a model to recognise satisfied and dissatisfied user states – as a first step towards a general model of the user's satisfaction state.

Keywords: human-computer interaction, emotion recognition, speech processing, user satisfaction

1. INTRODUCTION

If cognitive systems are to be used in everyday life, they must provide a pleasant and enjoyable user experience. This is especially important for companion-like cognitive systems: i.e. systems that adapt their functionality and behaviour according to the user's state, including his or her situation, abilities and requirements (cf. Biundo and Wendemuth (2016)). One of the key aspects here is user satisfaction, which is defined by Kelly (2009) as *"fulfillment of a specified desire or goal"* that the user wants to reach. Reliable automatic user satisfaction recognition would be useful for a broad variety of HCI applications, such as assistance for workers with different health-related requirements described by Vox et al. (2016) or robot-assisted physiotherapy described by Bächler et al. (2015). The first step towards such a model is the acquisition of useful, labelled sensor data for training and modelling. One challenge is to obtain the ground truth on this matter – the question is how to measure and label user satisfaction, since there is often no "correct answer" that is accessible to labellers, as satisfaction is a deeply subjective matter (cf. Kiseleva et al. (2016)). The obvious solution is to ask the user directly, e.g. by using questionnaires, as proposed by Shriberg et al. (1992) and Jekosch (2005). Such approaches are not always possible – for example, if the users are children as described by Kotzyba et al. (2015). But even if this is not the case, questionnaires are still not applicable to real-world scenarios, since they require interrupting the interaction. Another solution is to use labels of *complex social emotions* like frustration and empathy (cf. Chowdhury et al. (2016)) or implicit measures – e.g. the success of the task that the users try to accomplish (cf. Fox et al. (2005)). But here another problem arises, since acquiring this information for feature extraction might require offline processing and is not always possible in real time. One further source of information on user satisfaction used in various approaches is user behaviour – like scrolling and gazing behaviour (cf. Lagun et al. (2014)), and click-through data (cf. Joachims et al. (2005)). This data can be acquired directly, but the labelling process of such data poses the same problems as in the case of using social emotions.

In the investigation presented here, a different approach is chosen: The used data was collected in a Wizard-of-Oz (WOZ) experiment where users were asked directly whether they are satisfied – but not with the interaction as a whole, but with their own performance during the task at hand. This differs from the previously mentioned approaches, since in this case, the users speak about their satisfaction regarding themselves and not regarding the system. The most interesting aspect of this approach is that it allows a one-to-one mapping of the user utterance and his or her satisfaction level. This data can be used for developing a model for a satisfied user state, being a first step towards a general model of user satisfaction.

For this investigation, we develop two hypotheses. The first hypothesis H_1 is that it is possible to automatically distinguish between satisfied and unsatisfied user states using only acoustical features of user statements. The second hypothesis is aimed at the question of how the automatic classification can be improved and consists of two parts. Our hypothesis H_{2a} states that the results of this classification depend on which parts of the utterances are processed: For this, the model built only on voiced frames, e.g. frames containing only the voice of the user, is compared to the model built on whole user utterances containing pauses, noise and silence. Here, the expected outcome is that deleting the unvoiced frames leads to better performance. Our hypothesis H_{2b} states that the results also depend on the features that are used for the classification, and can be improved by employing a feature selection routine to reduce the dimensionality of the data. Here, the expected outcome is that using a reduced feature set achieves at least the same performance as using the full feature set.

The remainder of the paper is organised as described below. First, the state of the art regarding user satisfaction modelling is discussed. In Section 2, the data set used for the experiments is described. In Section 3, the developed labelling scheme is introduced, and the distribution of classes among different groups of users is presented. The experimental setup is explained in Section 4, including the two employed feature extraction routines, the feature set and the conducted experiments. The results regarding both working hypotheses are presented and discussed in Section 5. Finally, the investigation is concluded in Section 6, and some possibilities for further research are shown.

1.1 Related Work

Various investigations on the matter of user satisfaction, its measuring, detection and prediction can be found in the literature. User satisfaction is important in several fields, for HCI (e.g. for search applications cf. Hassan and White (2013), instant mobile messaging cf. Ogara et al. (2014), etc.) as well as other areas (e.g. for data warehouses like in the case of Soliman et al. (2000)).

Regarding results on automatic user satisfaction detection in the area of HCI, which is most relevant to the investigations presented here, several approaches can be found in the literature. Chowdhury et al. (2016) consider positive, negative and neutral user satisfaction obtained via measuring social emotions like frustration and empathy in human-human interaction (HHI) experiments, and achieve an F-measure of 40% - 61% for the classification of the satisfaction levels by using acoustic features alongside other, relatively sophisticated lexical and turn-taking features. Kiseleva et al. (2016) focus on search dialogues and use a self-assessment of users to investigate which interaction signals can be used to predict user satisfaction. They compare click, acoustic and touch features, achieving 61% - 79% F-measure. Feild et al. (2010) investigate the opposite of user satisfaction – *user frustration.* In their study, they develop a frustration model for users of a search application based on data collected by a mental state camera, a pressure sensitive mouse and a pressure sensitive chair, and further interaction features like query log (e.g. length of the query), scrolling behaviour and task duration. Depending on the data used, they achieve 54%-75% accuracy and 49%-80% F-measure for their predictions.

In the approach presented here, the goal is to investigate user satisfaction in the area of cognitive, companion-like systems and spoken interaction. For this purpose, well established, robust and easily obtainable acoustic features and user satisfaction data collected in WoZ-experiments are employed.

2. USED DATA - THE LAST MINUTE CORPUS

For the experiments presented here, the acoustic recordings of the multitmodal LAST MINUTE Corpus (LMC) (cf. Rösner et al. (2012a); Prylipko et al. (2014)) are used – the same recordings were object of examination for a number of times, regarding affective state recognition (cf. Frommer et al. (2012), Egorow and Wendemuth (2016)) as well as linguistic turns (cf. Rösner et al. (2012b) and Siegert et al. (2014b)).

The LMC contains 133 multi-modal recordings of German speaking participants (balanced regarding sex and age) during WOZ-experiments, where the participants are asked to accomplish close-to-real-life tasks with different levels of complexity. The setup revolves around a journey to the unknown place "Waiuku", which the participants have won. Each experiment takes about 30 minutes. Using voice commands, the participants have to prepare the journey, pack the suitcase, and select clothing. Most of the experiments are transliterated, which enables the automatic extraction of certain speaker utterances.

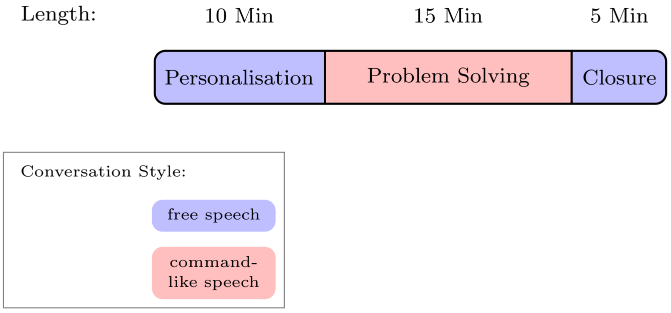

Fig. 1. Experimental design of the LMC, indicating the three different models and the two different conversation styles.

The WOZ-experiments consist of three modules with two different conversation styles: a personalisation module (free speech), a problem solving module (command-like speech) and a closure module (free speech). An overview of the experiment is depicted in Figure 1. The personalisation module, being the first part of the experiment, has the purpose of making the users familiar with the system to make their behaviour more natural. In this introduction stage, the users are encouraged to talk freely. During the subsequent problem solving module, the users are asked to pack a suitcase for their imaginary journey. The dialogue follows a specific structure, where the users perform certain actions (e.g. choosing an item) and the system confirms these actions. This part of the conversation is task-focused, therefore the users talk in a command-like style, which leads to a more or less strictly regularised interaction. The sequence of these repetitive dialogues is interrupted by pre-defined barriers for all users at specific points of the dialogue. These barriers are intended to increase the stress level of the users. Details on the design of the barriers can be found in Prylipko et al. (2014). Each experiment is completed with a closure module, where the system asks further questions about the task in general and users' satisfaction with their solution of the task, encouraging them to talk freely. After the experiments,

Table 1. Examples of instances of the five satisfaction levels.

S+ (very satisfied)	“very satisfied”, “extremely satisfied”, “absolutely satisfied”
S (satisfied)	“satisfied”, “quite okay”, “fairly satisfied”
M (moderately satisfied)	“moderately satisfied”, “so-so”, “reasonable”
U (unsatisfied)	“unsatisfied”, “not very satisfied”, “less satisfied”
U+ (very unsatisfied)	“very unsatisfied”, “absolutely not satisfied”, “not satisfied at all”

the participants could reflect their experience in semi-structured interviews, in which they indicated, that they experienced the system they interacted with as a hybrid system with partly technical and partly human abilities – a companion-like system (cf. Krüger et al. (2015), Krüger et al. (2016)).

3. DATA LABELLING

For the investigation presented here, the last part of the LMC experiments was chosen as data source. In this so called closure module, the users were asked about their satisfaction level. It should be emphasized that they were asked about their satisfaction regarding their own performance and not regarding the system. This data was chosen purposefully, assuming that the answer to this direct question represents a one-to-one-mapping between the user satisfaction state and the expressed vocal characteristics in his or her utterance.

As mentioned earlier, the first hypothesis of this investigation was that an automatic classification of the satisfaction level is possible using only the acoustic features of these answers. To label the data for this task, a labelling scheme based on the content of the answers was developed. This labelling scheme consists of five classes, namely the positive classes “very satisfied”, “satisfied” and “moderately satisfied”, and the negative classes “unsatisfied” and “very unsatisfied”. Some examples for the users’ expressions in each of the classes are presented in Table 1.

For the experiments described here, a subset of 89 out of a total number of the 133 LMC recordings was chosen for audio quality reasons. These 89 recordings were then annotated corresponding to the scheme described above. After the exclusion of speakers with unclear or missing answers on their satisfaction, a sub-set of 79 statements remained. Since the instances are not equally distributed over the five classes and some of the classes contain only few instances, it was decided to condense them into two nearly balanced classes – a positive class P, ranging from “very satisfied” to “moderately satisfied” and containing 37 instances, and a negative class N, including the categories “unsatisfied” and “very unsatisfied” and containing the remaining 42 instances. The detailed distribution of classes over these answers, as well as their aggregation into a positive and a negative class, can be seen in Table 2.

Table 2. Distribution of classes over the 79 participants: fine-granulated (Label$_f$) and binary granulated (Label$_b$)

Label$_f$	U+	U	M	S	S+
#instances	5	37	14	19	4
Label$_b$	N		P		
#instances	42		37		

Interestingly, the positive and negative classes are equally distributed among female participants as well as young participants, but there are more negative instances among male participants as well as among elderly participants, as can be seen in Table 3.

Table 3. Distribution of classes with respect to sex and age of the participants.

	P	N
Sex		
– male	14	21
– female	23	21
Age		
– young	22	20
– elderly	15	22

4. CLASSIFICATION EXPERIMENTS

The classification experiments presented here were performed using three different data pre-processing setups. In the first setup, the feature extraction took place on the users’ whole utterances, containing noise and pauses. In the second setup, a voice activity detection procedure was introduced as a pre-processing step to use only voiced parts of the users’ utterances for the feature extraction. In the third setup, again whole utterances were processed, but with a reduced feature set, which was selected using a random forest-based feature importance score.

4.1 Feature extraction

For the task of feature extraction, we decided to use the well-known *emobase* feature set provided alongside the openSMILE toolkit developed by Eyben et al. (2010), since it proved to be useful in a variety of affect-related investigations, on the LMC data employed here as well as on other data (cf. Tickle et al. (2013); Pfister and Robinson (2010)). The 988 features based on 19 functionals of 54 low-level descriptors (LLDs) were extracted on utterance level. Since each user uttered only one statement regarding his or her satisfaction level, this resulted in exactly one feature vector per user. This setup allowed us to implement a user-independent evaluation in a simple way, with each data instance corresponding to one feature vector and thus one user. After dividing the obtained data in training, development and test set, we standardised each set separately to zero mean and univariance ($\mu = 0, \sigma = 1$).

4.2 Voice Activity Detection

As mentioned above, we extracted the features from the user’s whole utterance. But such user utterances can contain filled as well as silent pauses, other sounds like knocking etc., and noise. Therefore, we developed a second hypothesis for our investigation, as already mentioned in Section 1: we assumed that such artefacts have an influence on the satisfaction recognition process. Our expectation was that they impair the classification results. To investigate this question, we employed a voice activity detection procedure to ensure that only frames containing voice were taken into account for feature extraction. For this task, we used the voice activity detection tool developed by Eyben et al. (2013) for the openSMILE toolkit. An example of the voice activity detection routine is shown in Figure 2. Here, the original sentence was *"ah hm ja geht so"* ("uh hm yeah so-so"). The voice activity detection found no voiced parts in the beginning of the utterance, and returned just the segment containing *"ja geht so"* ("yeah so-so") as voiced segments of the original sentence. For all of our data, 23% of the utterances included unvoiced segments which were deleted by the voice activity detection routine.

After obtaining only voiced segments of an utterance, we applied the feature extraction and the standardisation routines explained above, resulting in a second set of data.

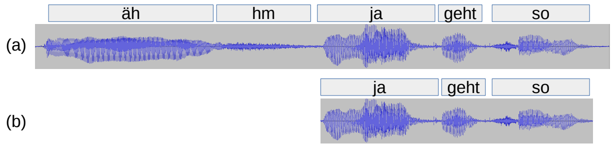

Fig. 2. An example for applying the voice activity detection routine: (a) shows the original statement *("uh hm yeah so-so")*, (b) shows the statement with deleted unvoiced frames *("yeah so-so")*.

4.3 Feature selection

Although the *emobase* feature set is widely used for a variety of emotion recognition investigations, it has one drawback: it has a relatively big number of features, namely 988, which might be too many for cases with only a limited number of data instances. To investigate this, we applied a common feature selection routine based on Random Forest (RF). RF is an ensemble learning method developed by Breiman (2001). The classifier consists of a number of decision trees, where in every node, a split feature is chosen from a randomly sampled subset of all the features – this split feature is the one that splits the data into classes according to the labels in the best possible way based on the impurity measure. Typically, measures like the Gini index or information gain are used for this. The RF training process can be used for feature selection, choosing the features that decrease the impurity most. To obtain the feature importance, a so called feature score is acquired by calculating how often and at which level each feature was selected in the individual trees of a trained RF. This score is then normalised by subtracting a random score – for this, the same procedure is run with shuffled target values in order to obtain the random bias of the data. The obtained difference indicates the importance of the feature for the data set. The procedure is described in detail by Chen and Lin (2006) and Silipo et al. (2014) and was successfully applied in a variety of investigations, e.g. by Menze et al. (2009), Statnikov et al. (2008) and Granitto et al. (2006).

For the investigation presented here, the feature selection procedure was applied to the training data set as described above. The obtained features were then ranked according to their score on the training data set. The top-most 10% of the features (99) were selected as feature set FS_{RF}. This value was chosen after several evaluation rounds on the development data set. The reduced feature set is based on functionals of a subset of 33 out of the originally 54 LLDs, such as certain MFCCs (e.g. MFCC[1] and MFCC[7]) and certain LSP frequencies, and their derivatives. There is also a remarkable amount of 21 LLDs that were not chosen in the feature selection procedure, such as F_0 envelope (F_0_env) and intensity. Figure 3 shows the selected LLDs according to the frequency of their occurrences in the reduced feature set.

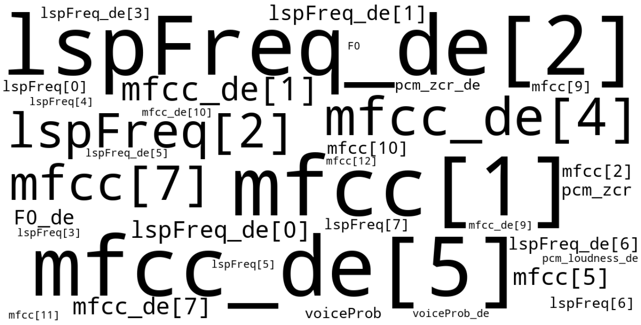

Fig. 3. Word cloud of the selected features. For reproducibility, the feature names supplied by openSMILE are used.

4.4 Used classifier

For classification, Support Vector Machines (SVMs) as described by Cortes and Vapnik (1995) were used. SVMs are widely used in the area of emotion recognition, also for benchmarking (cf. Tarasov and Delany (2011)). We chose the libSVM implementation developed by Chang and Lin (2011), and tested linear, polynomial and radial basis function kernels with different parameters on the development set for all three classification procedures – with and without voice activity detection, and with the reduced feature set FS_{RF}. For the first two setups, the linear kernel with C = 10 performed best. For the third setup, the radial basis function kernel with $\gamma = 0.001$ and C = 10 performed best.

4.5 Validation Strategy

In order to avoid overfitting, a user-independent evaluation is needed: the data for training and testing the model must not contain data of the same users. In general, there are two ways to ensure this. The first way is a leaving-one-out evaluation, where the training process takes place on the data of all users but one – this last user is reserved for testing. In the case of n users, the training and testing process is then repeated n times, and the results are then averaged over the n obtained models. This procedure ensures a generalisable result, but has two important drawbacks: first, it results in n different models instead of one single model, and second, it complicates fine-tuning the model's parameters, since the fine-tuning procedure would also have to be repeated n times and would lead to n different models with n different sets of parameters. Therefore, it was decided to choose the second way, a leaving-one-speaker-group-out (LOSGO) evaluation. The data was divided into three user-independent subsets: a training set for training the model, a development set for fine-tuning the model's parameters and a test set for evaluating the model. The development and the test set contain statements of 8 users each, the training set contains statements of the remaining 63 users. The three sets have a nearly equal distribution of sex, age and classes, as shown in Table 4. Since the different data sets include data of different user groups regarding age and sex, this setup ensures generalisability.

Table 4. Distribution of sex, age and class over the training, development and test sets.

	Sex		Age		Class	
	F	M	Y	E	P	N
Training	35	28	33	30	30	33
Development	4	4	4	4	4	4
Test	4	4	4	4	4	4
Overall	44	35	42	37	37	42

According to the principle of LOSGO evaluation, the models were trained on the training set and the parameters of the classifier were fine-tuned on the development set. After choosing the best parameters, the model was evaluated on the test set. This procedure was conducted for all three setups: classification without voice activity detection, classification with voice activity detection, and classification using the reduced feature set obtained by feature selection.

4.6 Performance Measures

For the evaluation, the established performance measures recall, precision and their harmonic mean, F-measure, were used. The measures were first calculated separately for the two considered classes P and N, before averaging them over these two classes, resulting in an unweighted average value for recall, precision and F-measure.

5. RESULTS AND DISCUSSION

The results obtained in the conducted experiments are shown in Table 5 in terms of recall, precision and F-measure for each class as well as the unweighted average for all obtained results: classification without voice activity detection and with voice activity detection, and also the classification using the reduced feature set FS_{RF}.

5.1 Classification without voice activity detection

For the first data pre-processing setup – the one where the whole user utterances without deleting unvoiced frames were used – the classification achieved an unweighted average recall of 87.5% and an unweighted average precision of 90%. The results on the test set are equal to those on the development set, which shows that the classification did not overfit to the development set. Overall, we can say that the negative class N could be recognised better than the positive class P (recall of 100% vs. recall of 75%, respectively) – this is an interesting outcome, since in real-world applications, it is more important to detect dissatisfaction to be able to react to this user state appropriately.

Table 5. Classification results for the satisfied (P) and the dissatisfied (N) classes and their unweighted average (UA) for development as well for the test set: without voice activity detection (w/o VAD), with voice activity detection (with VAD) and with feature selection (with FS_{RF}). The unweighted average results on the test sets are highlighted in gray, the best results are printed in **bold**.

w/o VAD	Recall, %	Precision, %	F-Measure, %
Dev			
– P	75.0	100	85.7
– N	100	80.0	88.9
– UA	87.5	90.0	87.3
Test			
– P	75.0	100	85.7
– N	100	80.0	88.9
– UA	**87.5**	**90.0**	**87.3**
with VAD	Recall, %	Precision, %	F-Measure, %
Dev			
– P	50.0	100	66.7
– N	100	66.7	80.0
– UA	75.0	83.3	73.3
Test			
– P	50.0	100	66.7
– N	100	66.7	80.0
– UA	75.0	83.3	73.3
with FS_{RF}	Recall, %	Precision, %	F-Measure, %
Dev			
– P	75.0	100	85.7
– N	100	80.0	89.0
– UA	87.5	90.0	87.3
Test			
– P	75.0	60.0	66.7
– N	50.0	66.7	57.1
– UA	62.5	63.3	61.9

5.2 Classification with voice activity detection

Considering the results for the second data pre-processing setup – the one with deleted unvoiced frames – the classification did not perform as well as in the first condition. This finding contradicts to the working hypothesis H_{2a} assuming that pauses and noise impair the classification performance. On the contrary, by excluding such artefacts, an unweighted average recall of only 75% and an unweighted average precision of only 83.3% could be achieved, resulting in an F-measure decrease of 14% absolute. Again, here we can find that the class N could be recognised better than the class P (recall of 100% vs. recall of 50%, respectively), as in the previous case.

5.3 Classification with feature selection

For the third condition – using a reduced feature set previously selected using RF – the results on the development data set and the test data set differed remarkably. For the development data set, the same results were achieved as with the complete data set (i.e. the first condition). But on the test data set, both the unweighted average recall and the unweighted average precision dropped by 25% absolute and 26.7% absolute, respectively. This is mostly caused by the drop in recall for the class N (recall of 100% on the development data set vs. recall of only 50% on the test data set). This decrease in recall leads to a substantial drop in precision for both classes.

5.4 Discussion

The presented classification experiments show that it is possible to recognise high and low satisfaction from user utterances with a feasible recognition rate using only the acoustic content of the user statements on their satisfaction. This proves our first hypothesis H_1. The next finding here is that processing whole utterances, including silence and noise, does perform better than if removing such unvoiced frames before feature extraction. This contradicts to our hypothesis H_{2a}. From this insight, we conclude that besides the acoustics of the pure utterances, the paralinguistic content such as filled and silent pauses and discourse particles, is also an important source of information for this task and their modelling should not be neglected – similar to findings in Siegert et al. (2014b) and Lotz et al. (2016). Another interesting result is the effect of feature selection on the classification performance: after the application of a common feature reduction method using random forests, the classification performance dropped substantially. Our hypothesis H_{2b} was that by reducing the data dimensionality, the performance of the classification will either reach the same level or even improve – but this assumption was not backed by our experimental results. As already mentioned above, the selected feature set contains mostly MFCCs and LSP frequencies – in fact, 13 of the 99 features are related to the feature MFCC[1] and its derivative, 12 are related to the feature MFCC[5] and its derivative, summing up to over 25% of all the selected features. Features based on other LLDs – such as other MFCCs, voicing probability and intensity – occur rather rarely or even do not occur at all. This leads to the conclusion that even if a feature is not as important as the random forest-based method suggests, it still can carry information crucial to the classification.

The performance of the user satisfaction classification presented here is comparable to those of Chowdhury et al. (2016) and Kiseleva et al. (2016), mentioned earlier. In contrast to those approaches, we employ acoustic features that can be extracted automatically and which are not interaction-related or behavioural features. But the results can hardly be compared directly, since the data bases and events in the data are different. Chowdhury et al. use data of HHI, namely a call-center corpus, with annotations of complex social emotions like satisfaction, dissatisfaction, frustration and empathy. Kiseleva et al. develop their method on data of HCI gathered during search experiments, by questioning the users on their experience with the *system* (*"how satisfied are you with your experience in this task"*). The first approach needs detailed annotations done by expert annotators, since it is not an easy task to recognise spontaneous, complex emotions even for humans – it was shown by Scherer (1981) that for acoustic data without context, human annotators achieve an average accuracy of only 60%. One solution for this problem is aiming for a high inter-rater reliability, which demands a bigger amount of data (cf. Siegert et al. (2014a)). The second approach does not have this problem, since the satisfaction level is assessed directly by asking the user. This case is similar to our case, the only difference here is the subject of satisfaction – we considered the users' satisfaction regarding their own performance.

6. CONCLUSION

The presented investigation shows that it is possible to automatically classify the state of satisfaction from speech acoustics, without including the speech content. For this task, data acquired during spoken WoZ-interactions was used. In this data, the users were asked about their satisfaction with their own performance during a task. Using the emobase feature set and SVM for classification, we achieved an F-measure of 87.3% in a subject-independent two-class setting. Surprisingly, deleting unvoiced frames using a voice activity detection routine did not improve the results – on the contrary, this procedure lead to a decrease in recall and precision, and thus also in F-measure. Furthermore, feature reduction also did not improve the results.

There are three main questions that are left for further investigation. One important question is whether the model developed here can be used for other data, i.e. whether the state of user satisfaction in case of satisfaction with *the user's own performance* is comparable to satisfaction with *the system's performance*. The next question is, which and how many features are suited best for the task of user satisfaction prediction. The approaches representing the current state of the art in this regard use different kinds of features, ranging from acoustic ones as in our case to behavioural ones, like turn-taking. For real-world-scenarios, the obtainability of the features and the robustness of the feature extraction should be in the focus – in this context, acoustic features have proven to be easy to obtain in high-quality audio data, but their robustness against noise in "in the wild" scenarios should be further investigated. And finally, there is the *temporal* aspect of user satisfaction: *At which point* is it possible to say that the user is satisfied or dissatisfied with a dialogue? If user dissatisfaction is recognised early in the dialogue, the system can adapt itself to the user's state and act appropriately. These questions should be addressed in further research.

ACKNOWLEDGEMENTS

The authors thank for continued support by the SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (www.sfb-trr-62.de) funded by the German Research Foundation (DFG). Further, this work has been sponsored by the German Federal Ministry of Education and Research (BMBF) in the program Zwanzig20 Partnership for Innovation as part of the research alliance 3Dsensation (www.3d-sensation.de/) under grant number: 03ZZ0414.

REFERENCES

Bächler, L., Bächler, A., Kölz, M., Hörz, T., and Heidenreich, T. (2015). Über die Entwicklung eines prozeduralinteraktiven Assistenzsystems für leistungsgeminderte und-gewandelte Mitarbeiter in der manuellen Montage. *Kognitive Systeme*, 1.

Biundo, S. and Wendemuth, A. (2016). Companion-technology for cognitive technical systems. *KI-Künstliche Intelligenz*, 30(1), 71–75.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Chang, C.C. and Lin, C.J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 1–27.

Chen, Y.W. and Lin, C.J. (2006). *Combining SVMs with various feature selection strategies*, 315–324. Springer.

Chowdhury, S.A., Stepanov, E.A., and Riccardi, G. (2016). Predicting user satisfaction from turn-taking in spoken conversations. In *Proc. of INTERSPEECH'16*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273–297.

Egorow, O. and Wendemuth, A. (2016). Detection of challenging dialogue stages using acoustic signals and biosignals. In *Proceedings of 24th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 137–143.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of the ACM MM-2010*, 1459–1462.

Eyben, F., Weninger, F., Squartini, S., and Schuller, B. (2013). Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 483–487. IEEE.

Feild, H.A., Allan, J., and Jones, R. (2010). Predicting searcher frustration. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 34–41. ACM.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2), 147–168.

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

Granitto, P.M., Furlanello, C., Biasioli, F., and Gasperi, F. (2006). Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2), 83–90.

Hassan, A. and White, R.W. (2013). Personalized models of search satisfaction. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2009–2018. ACM.

Jekosch, U. (2005). *Voice and speech quality perception. Assessment and Evaluation.* Springer, Berlin.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, 154–161.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2), 1–224.

Kiseleva, J., Williams, K., Jiang, J., Awadallah, A., Zitouni, I., Crook, A., and Anastasakos, T. (2016). Predicting user satisfaction with intelligent assistants. In *Proc. of ACM SIGIR'16*, 495–505.

Kotzyba, M., Siegert, I., Gossen, T., Nürnberger, A., and Wendemuth, A. (2015). Exploratory voice-controlled search for young users : Challenges and potential benefits. *Kognitive Systeme*, 1.

Krüger, J., Wahl, M., and Frommer, J. (2015). Making the system a relational partner: Users' ascriptions in individualization-focused interactions with companion-systems. In *Proceedings of the 8th International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2015)*, 48–54.

Krüger, J., Wahl, M., and Frommer, J. (2016). Users relational ascriptions in user-companion interaction. In *Proceedings of the 18th International Conference on Human-Computer Interaction (HCI 2016)*, 128–137. Springer.

Lagun, D., Hsieh, C.H., Webster, D., and Navalpakkam, V. (2014). Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 113–122. ACM.

Lotz, A., Siegert, I., and Wendemuth, A. (2016). Comparison of different modeling techniques for robust prototype matching of speech pitch-contours. *Kognitive Systeme*, 1.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F.A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10(1), 213.

Ogara, S.O., Koh, C.E., and Prybutok, V.R. (2014). Investigating factors affecting social presence and user satisfaction with mobile instant messaging. *Computers in Human Behavior*, 36, 453–459.

Pfister, T. and Robinson, P. (2010). Speech emotion classification and public speaking skill assessment. In A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli (eds.), *Human Behavior Understanding*, volume 6219 of *LNCS*, 151–162. Springer Berlin Heidelberg.

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

Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., and Otto, M. (2012a). LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions. In *Proc. of LREC'12*, 96–103.

Rösner, D., Kunze, M., Otto, M., and Frommer, J. (2012b). Linguistic analyses of the LAST MINUTE corpus. In J. Jancsary (ed.), *Proceedings of KONVENS 2012*, 145–154. ÖGAI.

Scherer, K.R. (1981). Speech and emotional states. *Speech evaluation in psychiatry*, 189–220.

Shriberg, E., Wade, E., and Price, P. (1992). Human-machine problem solving using spoken language systems (sls): Factors affecting performance and user satisfaction. In *Proceedings of the workshop on Speech and Natural Language*, 49–54. Association for Computational Linguistics.

Siegert, I., Böck, R., and Wendemuth, A. (2014a). Interrater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. *Journal on Multimodal User Interfaces*, 8(1), 17–28.

Siegert, I., Prylipko, D., Hartmann, K., Böck, R., and Wendemuth, A. (2014b). Investigating the form-function-relation of the discourse particle hm in a naturalistic human-computer interaction. In S. Bassis and A.E. amd Francesco Carlo Morabito (eds.), *Recent Advances of Neural Network Models and Applications*, volume 26 of *Smart Innovation, Systems and Technologies*, 387–394. Springer.

Silipo, R., Adae, I., Hart, A., and Berthold, M. (2014). Seven techniques for dimensionality reduction. Technical report, KNIME.

Soliman, K.S., Mao, E., Frolick, M.N., et al. (2000). Measuring user satisfaction with data warehouses: an exploratory study. *Information & Management*, 37(3), 103–110.

Statnikov, A., Wang, L., and Aliferis, C.F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1), 319.

Tarasov, A. and Delany, S.J. (2011). Benchmarking classification models for emotion recognition in natural speech: A multi-corporal study. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 841–846. IEEE.

Tickle, A., Raghu, S., and Elshaw, M. (2013). Emotional recognition from the speech signal for a virtual education agent. *J. Phys.: Conf. Ser.*, 450, 012053.

Vox, J.P., Franz, S., and Wallhoff, F. (2016). Adaptive Bewegungsanalyse von physiotherapeutischen Übungen für eine optimierte Mensch-Roboter-Trainingsinteraktion. *Kognitive Systeme*, 1.