Toward a self-adapting resource-restricted voice-based Classification of Naturalistic Interaction Stages

Norman Weißkirchen (norman.weisskirchen@ovgu.de) *
Ronald Böck (ronald.boeck@ovgu.de) *,**

* *Otto-von-Guericke-University Magdeburg, Magdeburg, 39016 Germany*
** *Center for Behavioral Brain Sciences, Otto-von-Guericke-University Magdeburg, Magdeburg, 39016 Germany*

Abstract: For the implementation of a user oriented assistance system, or an advanced human-computer interface, the computerized part needs to be supplied with a wide range of information concerning the current user. The adaption and personalization of the user experience is based not only on the recognition of control commands, which can be solved through current voice controlled interfaces, which are capable of interpreting the syntax of its user, but also on the non-syntactical information conveyed through voice affect or body language. Therefore, current research aims to advance the capabilities of computer assisted systems to recognize emotions and also more general states of a user. This is done to provide for the anticipative and cooperative behavior needed for assisting applications, which then can adapt to the specific needs the user can have concerning their personal state and level of mental involvement. To facilitate this, one needs to provide current classifiers working on highly relevant features suitable for the discrimination of user states. Used features can change considerably between different classification tasks which hinders a general set of features. At the same time such systems often need to be mobile applications, which restricts the amount of computational power available, and additionally real-time compatible, which limits the complexity of the used classifier and its applicable feature set, since resource-restricted hardware is used in such application. Our aim is to provide an alternative approach to get similar results to more complex methods, by using simpler architectures with a relatively small dataset. The inspiration were comparable studies concerning the minimization of the used feature set as discussed in this paper.

Keywords: machine learning, feature set, classifiers

1. INTRODUCTION

Voice controlled human-machine interactions are an important part of current research. The underlying idea of a non-haptical controlled assistance system contains the aim for a natural and easy way to interact with a machine. This is important for a wide variety of systems, for example in the service area or as an assistant tool for handicapped people. As not everyone can or want to learn complex control mechanisms, voice control provides a way of interaction close to the standard human method of interaction. This is also helpful for first time users or people not used to technical systems.

Another important aspect of the current human-machine interaction is the provision of additional information about the current user. As machine do not have empathy yet, interacting can feel again lifeless and in certain situations unforgiving, which in turn leads to animosity from the user towards the system originally planned to support the human. To alleviate this problem a machine needs to know what the general situation of its user is and adapt its way of interacting based on this information. An angry or overworked user, for example, would be treated differently than a happy and energetic one, as the underlying system would adapt and try to extract the reason for the dissatisfaction and, if possible lessen or remove it.

The general method to provide this information is by employing a voiced based classifier, which filters the audio recording of a user for intent, emotion or general state of mind (Nass et al. (1994)). This can also be used for anticipative reactions from the side of the machine as certain reactions are a response to certain human mental states. These classifiers are often from the area of artificial neuronal networks, to employ the advantages of machine learning methods, which helps finding the complex relations between the measurable features and the corresponding information which can be inferred from them. Additional methods include the addition of further sensors, like cameras, to provide a multi-modal array and to enhance the confidence of the classifier.

⋆ funded by the European Funds for Regional Development (EFRE) and by the Federal State of Sachsen-Anhalt, Germany, under the grant number ZS/2017/10/88785

In this paper, the main focus was on voice input as information channel (Biundo and Wendemuth (2016)), as this is more concerning to the experiment at hand. Otherwise, for example in real world applications, the use of cameras can be understood as a great intrusion into the privacy of a user, which in turn can make people uncomfortable, marring the results of the classification.

One of the problems in combination with voice based recognition, is the search or decision for the used feature set, as the work with classifiers has shown that different tasks tend to favor different feature sets. The solution of using all available features, does not automatically bring the best results as the amount of data leads to curse of dimensionality problems and information overlapping which generates complexity for the whole system compared to a more specialized architecture capable of classifying only one or two situations. Another point is the difference of stable and robust classifier and more specialized ones. Humans tend to adapt to the way people are speaking not only in the semantic way but also concerning to the tone of voice, for example the expectation how a happy young man sound would be different from a happy older woman (de Looze and Rauzy (2011)), as well as the specific ways a human communicates, based on upbringing or regional influences. The same adaption is not the normal way for machine learning applications which search for general features differences, which are the same over all groups of users.

The method to improve on that was chosen to be a two staged classification approach, the first stage is a general classification based more on robustness which is able to separate the speakers into certain groups which tend to have the same tone of voice and comparable reactions to each other. The next step is a more specialized classifier with a higher efficiency for this smaller subgroup. Comparable experiments have shown that by concentrating on specific features for groups of speaker one can improve the classification results (Böck et al. (2018), Siegert et al. (2018)). By adding this to systems after each other, better results were achieved by employing different feature sets, while at the same time reducing the number of features used for each step. As each feature used in a classifier need exponential more computational power this is preferable to classifier approaches using full feature sets containing hundreds to thousands of features.

An additional area for this research is the use of this system as a mobile application, as shown by porting this system to a raspberry pi 3. One of the aims is to provide this kind of phased approach for a just in time usage, in opposition to more complex approaches requiring excessive computational power. As mentioned in the abstract it is also in comparison to similar studies towards the reduction of necessary features, as for example done for the Geneva minimalistic acoustic parameter set (GeMAPs) (Eyben et al. (2016)) and for the Last Minute Corpus (LMC) (Rösner et al. (2012)).

1.1 Research Questions

The aim is to efficiently classify the current state of a human user in regard to its mental load, or state of mind. One can distinguish between low mental load caused by repetitive and easy interactions, and high mental load, appearing when a user is under stress caused by complicated work tasks or aggravated by the situation at hand. This translates to a general level of involvement from the user, which in turn can be used to determine if lacking attention can be attributed to reaching the limits of the cognitive resources of the user or the user getting disinterested in the situation at hand. This distinction influences the way an assisting system has to react to maximize the help it can provide. An important point of this experiment is the authenticity of the used data set, which is non-acted and natural, and the general approach to the problem which follows a roughly real world application. The system implemented is designed to work on a mobile framework and gets result in real-time spans, without the need for external support. The general ability to classify emotions or human states are already being done in comparable work (Picard (1999)) and is as such not the main consideration of this paper.

The main points and questions examined are:
Q1. Is the optimal feature set the same for all speakers?
Q2. If not, can one divide and/or adapt the classification task in such a way as to reduce complexity?
Q3. How can one use this distinction in a mobile classifier assembly, while at the same time trying to reduce the necessary overhead?

1.2 Related Work

The research is based on a variety of other works. The general application of using voice as an input for affect classification on humans is one of the current pillars of human-machine interaction (Picard (1999), Ward and Marsden (2004)). This represents a state of the art approach for naturalistic and efficient dialogs between users and assistance systems. As this system is primarily used for direct support, it needs to be capable of real-time classification of naturalistic utterances. As such the shift of used datasets changes from studio quality recordings of actors (Burkhardt et al. (2005)) to more realistic, unannounced recordings (Griffiths and Scarantino (2009)), which can also contain noise and non-ideal preparation of the audio before further processing.

One of the main points of this research is the inability to use the same features for a wide variety of different tasks. While the amount of extractable features from voice is quite high, especially in the context of classification tasks (Eyben et al. (2016), Scherer (2001), Tahon and Devillers (2016)), and at the same time the trend for more complex architecture is apparent, one can face a sometimes exponential rise of complexity and combined with that longer recognition times. This is in contrast to the trend of mobile applications as used in smartphones and smart assistance. The current way of solving is often the exportation of complex task to cloud based solutions, which necessitates not only high band-with of data transfer capabilities but also poses a problem in the future when more appliances should be connected to networks e.g. Internet of Things, as well as problems concerning the safety of user data, which has to be transmitted and stored externally. The aim is instead to reduce the amount of

Fig. 1. View of an interacting user with the voice controlled system from Tornow et al. (2016), while the way the people sat in front of the microphone changed during the course and between different user, the general structure remained the same

needed features, while at the same time using simpler classifier, which are instead build in line and used as needed. This has been done partial for acted emotions (Frommer et al. (2012a), Siegert et al. (2014), Egorow and Wendemuth (2016)).

2. CORPUS

For the experiments in this paper the used dataset was the integrated Health and Fitness (iGF) - Corpus, for which further information can be found in (Tornow et al. (2016)). As a general outline, this is a dataset which was conceived for the multimodal measurement of mental workloads of elderly people while performing various tasks. The set originally contains information from multiple sensorial inputs, for example audio and video recordings, as well as body posture or bio-physiology (e.g. heart rate, blood pressure) for a full amount of ten input channels. For the experiment this amount was reduced to one audio recording taken by a stationary shotgun microphone. With this one can simulate an application in which the user would consult an independent system, which should recognize the current (mental) state of the user. Further inputs could potentially help in this distinction, but would detract from the original aim of the research on a mobile and adaptable system. Further modalities (except for the video recording) would imply a knowledge of the current user and a preparation phase in which the system could already been adapted to the specifics of the speaker. By using the external microphone, the system has to be capable of working with slightly different volume levels (as produced by the way the user is sitting in front of the microphone). At the same time the general situation and relation of objects is the same for all the participants, as dictated by the placing of the different parts of the system (desk with microphone and monitor with chair in front of it), as can be seen in Fig. 1.

The recordings in this corpus are taken from 65 participants, from which there are 45 female speakers and 20 male ones. As said before the focus was on elder people with ages from 50 to 80 years old. The average over all speakers is 66 years while the average for male participants is slightly above this point with 68 years, while the average age of the female participants is slightly lower at 65 years of age. The corpus was recorded without the direct knowledge of the participants concerning the aim of distinguishing the different mental states from each other. The pretense of the recording was a medical examination, concerning the variation in different kinds of gaits when elderly people, with and without former medical problems in their movements, partake in a moderately demanding test course. Specifically, the audio recording was taken from the interaction between the participants and the (supervised) machine controlling and directing the tests.

The test conducted by the machine was divided in turn in five stages which were also conveyed to the participants under slightly different pretext. The order of the stages and the content of the test within were planned to lead the participants into experiencing different dispositions and mental states (Rösner et al. (2012), Frommer et al. (2012b)). While the induction of this stages was not conveyed to the users, the staged approach with evolving difficulty level fits also natural in the perception of this kind of physical test. With this situation one can exclude any acting of the mental state from the point of view from the participants in excess of the normal reactions of the users concerning the situation they were in. As the machine only conversed trough written text and a monotonous female machine voice, one can also remove any subconscious influence from the test supervisors.

To explain the different stages in a little more detail, here first the graphical representation of the original testing plan of the experiment with the several sub-stages in Fig.2. Of note is, that after each stage the user has time to relax for several minutes while music is playing. Together with the stage wise addition of more stressful situations, this was done to generate a more neutral starting situation at the beginning of each stage, so that the influence of former stages is minimized (Panning et al. (2012)).

The Corpus begins with two stages with comparatively neutral mental state of the user. This two stages "introduction" and "interest" are mainly done for describing to the user the backstory of the test, together with a general explanation of the usage of the assistant system and the available task. The user learns to interact by a dialogue system, in which the user and the machine interact by voicing their questions and responding answers. The user state is mostly neutral, with a slight involvement, as the questions of the system were answered.

One of the main stages of interest is the third, the "underload" stage. Here the user has to repeat given texts concerning the definition and work of the different tasks several times. This repetition, without outer time constrains, allows for a more relaxed state of mind. This is done to achieve a bored and/or low involvement in the user. The recorded speech is either the text of the manuals or the general control commands, e.g. "Proceed" or similar utterances. This stage is the source for one state for the trained classifier.

The other main stage for this classifier is the fourth stage, the overload stage. The user gets complex instructions while at additionally having a time constrain. Furthermore, there is a review phase in which the user, get confronted with nonexistent errors of former stages. This brings the user in an agitated state, in which he or she

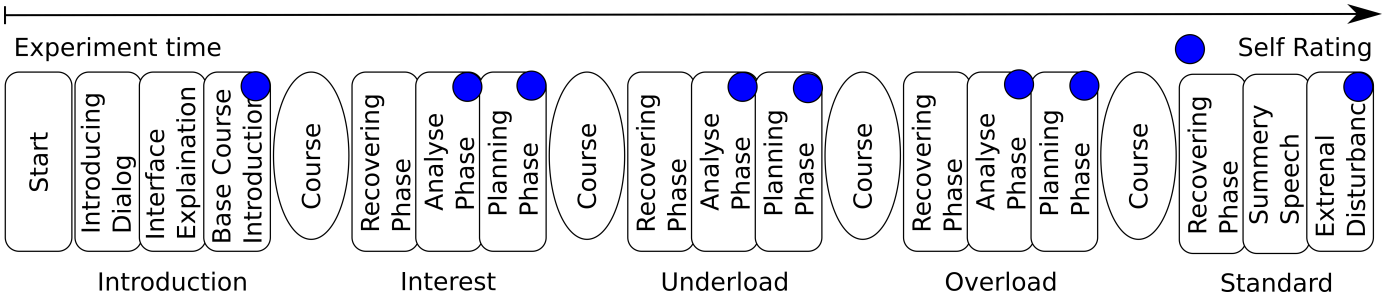

Fig. 2. Plan of the iGF experiment as taken from Tornow et al. (2016), shown are the different stages of the experiment with their internal thematic partitions and the external self-rating with which the emotion/affect of the subject was determined.

is either angry at the system and/or actively trying to solve the problems. Generally, the user exhibits a strong interaction with the system. For this state it is not relevant if the user beliefs the system or not, as the involvement is stronger in both situations. This is the main contrast to the former stage, which were classified in the experiments conducted.

The last stage uses unannounced alarms for short burst of emotion, because of its relative shortness, and the lack of voiced reaction from the users, this stage was not used in the experiment.

3. EXPERIMENTAL SETUP

The recordings were splitted into their respective users and the different stages. Before the extraction of the features, all traces of the machine speaking were removed (which can be easily done in a real environment by simply removing the audio track of the machine from a recording) and separated the audio in smaller utterances automatically by cutting the recording at all pauses longer than three seconds. All utterances below one second were removed, which removed nearly all noise and echoes as well as impulses caught in the recording. A certain amount of noise is expected and wanted, to compare the situation to real world appliances in similar situations. The remaining utterances build the basis for the following experiments.

The feature set used was the *emoBase* feature set as generated by the openSMILE Toolbox (Eyben et al. (2013)). This set contains 988 prosodic and spectral features derived from 19 functionals calculated for 52 low level descriptors (LLD) and is one of the standard sets used for emotion and affect recognition. It contains Mel Frequency Cepstral Coefficients (MFCCs) which remain as one of the state of the art features for classifier tasks (Schuller et al. (2009), Böck et al. (2010), Schuller et al. (2011)), Linear Spectral Pairs (LSP) (Shahzadi et al. (2013)) or intensity and loudness. The full set is built from the fundamental frequency (F0), envelope of the fundamental frequency contour (F0env), LSP 0 to 7, MFCC 0 to 12, intensity, loudness, zero crossing rate (ZCR), and voicing probability (voiceProb) as static descriptors together with the delta coefficients of these static descriptors (cf. Table 1).

As the full set contains 988 features, the first experiment was concerned with the possibility to reduce this amount without losing the ability of the classifier to distinguish the stages.

Table 1. The number of used features in the *emoBase* feature set, separated according to measured aspect and between measured and derivate feature

Category	Features + Derivates
PCM Intensity	19 + 19
PCM Loudness	19 + 19
MFCC (in 12 Groups)	228 + 228
LSP Freq (in 8 Groups)	152 + 152
PCM ZCR	19 + 19
Voiceprob	19 + 19
F0	19 + 19
F0 Env	19 + 19
Overall	494 + 494 = 988

3.1 Distinction of Features

To find the features most useful for the distinction of the current mental load phases of the user the examination concentrated on different occurrences. The first was an examination of the differences general to all utterances when they were taken from the underload stage and when they were taken from the overload stage. For this the average feature value for each speaker was taken in each stage and make a comparison using the Kruskal-Wallis test (Kruskal and Wallis (1952)). For this the truth of all feature values belonging to the same population were determined. The result for this experiment were taken though the p-value, ranging from 0.05 for significant belonging to different populations to values below 0.001 for most significant belonging to different populations. As the finding of one speaker is not conclusive for examination this was examined over all speakers of the test. As border for the experiment the distinction was set at roughly two-third of the whole speaker group as representative. This means when the feature values where more than significant different between the two stages for more than 66% of the speakers this was taken as an indicator.

In a second step the difference of values extracted between the stages were taken and a correlation analysis concerning the different subgroups of the speakers were conducted. For this the changes happening during the change from underload to overload were looked at, specifically if it is a positive or a negative change. For a more specific approach the corpus was separated in male and female. This was

used for a further correlation analysis with the other speakers, now in conjunction with their sex and rough age grouping. As the general voice of male and female are different from each other but relatively similar within each group, one can assume a certain correlation inside these groups. An examination of the general age grouping was also performed. While all speakers are above 50 years of age, only a rough distinction could be established between the speakers closer to 80 and the ones closer to 50. As there is a certain way voices changes with age, one can presume bigger correlations in the respective age ranges (Brown et al. (1991)).

For the comparison between the groups the Spearman correlation coefficient was used. This allows for an analysis when the data is not normally distributed (Spearman (1904)). The result could be either a positive correlation, negative correlation or no correlation. With this, similarities inside the subgroups could be found.

3.2 Mobile Application

The second part of this experiment concerned itself with the possibility to replace a given network with a high amount of available features with smaller ones, only using a much smaller set of features. For this the corpus was used as explained in the first part of the experiment. A cross-validation was performed to test the result of this classification.

For this experiments Random Forests (RF) and Support Vector Machines (SVM) as classifier were used. The RF had no maximum depth for the base examination and were stopped at a depth of 10 for the reduced feature set. The amount of trees was set to 100. For the SVM a linear kernel was used. In this case one has a two-class classifier, underload and overload, and a three two-class classifier for the staged approach (one for the distinction of the sex and two for the smaller groups again with underload and overload).

For further experiments the training set was reduced by changing the classifier task. Instead of finding underload and overload for all speakers only all the male or all the female speakers were looked at. To reduce the base group from which one has to distinguish the different speakers, different feature sets were also tested. The Full Set is taken as comparison, the smaller set were taken from the solutions found in the first experiment. As one of the aims is the implementation on mobile systems, as the raspberry pie 3, one aim was to reduce the complexity and necessary data storage on any system. As the available storage and processing power is limited the aim was to fit into these parameters.

4. RESULTS

The first point is the search for relevant features in combination for the mental load of the user, while the second part is the test for an implementation of a staged classifier.

4.1 Important Features

The search was for features which not only provide a significant difference between the two main stages, underload and overload of the mental state, but also provide a comparable correlation between different speakers. As correlation is in the range from -1 to 1 one can average the values over certain groups and become so a result on how homogenous these groups are in their features. When looking at the general correlation between all speakers with the full feature set, around 0.3 of average correlation was measured. Additionally, while comparing inside the sex groups only neglectable better results of 0.31 and 0.32 for female and male speakers were achieved. This low value presents the inherent advantages of machine learning solutions which can learn which input features get weighted more. When looking at the correlation between relative and absolute value differences even lower values were measured.

Table 2. Features sorted by the number of times they were significantly different between underload and overload stages (sma - smothed by moving average filter, de - derivate, amean - arithmetic mean of the contour, linregc1 - slope of a linear approximation of the contour, range - max-min of values, maxPos - frame position of the maximum Value, zcr - zero crossing rate)

Name	**Times of Features are: Most/Higly/Just Significant**
pcm_intensity_sma_de_amean	50 / 57 / 59
pcm_loudness_sma_de_amean	48 / 52 / 57
pcm_intensity_sma_de_linregc1	47 / 51 / 56
lspFreq_sma2_range	47 / 53 / 56
lspFreq_sma1_range	47 / 54 / 55
mfcc_sma_de1_range	45 / 53 / 59
lspFreq_sma_de0_maxPos	44 / 52 / 56
pcm_zcr_sma_de_maxPos	43 / 52 / 58
pcm_zcr_sma_de_range	43 / 49 / 58
pcm_zcr_sma_range	42 / 51 / 59

As said in section 3.1 a Kruskal-Wallis calculation to search for features was used. The idea behind this is, that features which values are significant different, or which distribution suggest a different population, provide more information about the mental load, than features in roughly the same distribution. Looking at the different p-values and the amount of speakers in which they are significant different and searching for occurrences which are repeated often enough to not amount only for alpha inflation, as given by the numbers of examples. Concentrating on the top ten features as shown in Table 2 which are also the one which are most significant for more than two-third of the speakers.

Considering the different sex groups further observations are on hand. The difference between male and female speaker concerning the features is primarily given by the relative amount of speaker which have significant differences. This means that from the male speakers there is a nearly homogenous significance over all speakers, as 19 of 20 of the speakers have the same relevant features, which means 95%. On the other hand, the female speakers have much lower amount of comparable number of speakers with the same reaction with only 32 of 45 female speakers sharing the same features, which means 71%.

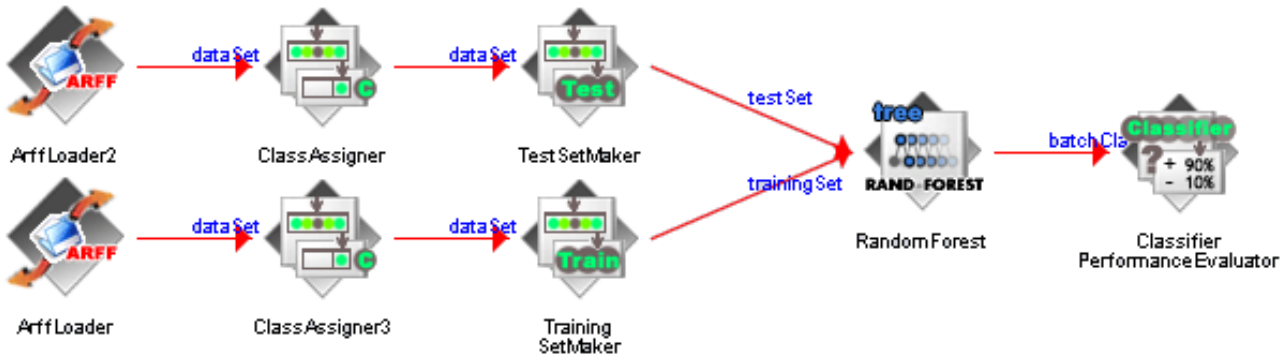

Fig. 3. The architecture is depicted on as implemented in WEKA (Frank et al. (2016)), the test-set (User) and train-set (other speakers) and the result can be taken from the generated output file, for SVMs the classifier was replaced.

Table 3. Correlation of the top 10 features, after reducing the set to the most significant features

	male	female	both
male	0.89	0.87	0.87
female	0.87	0.87	0.87
both	0.87	0.87	0.87

Looking at the correlations reduced on these top 10 features considerably better results were achieved. As shown in Table 3 around 0.87 of average correlation with better results were measured when only looking at the male speakers. Often there were correlation in excess of 0.95 for many speaker, with the most detracting groups of speakers, in the sense of the speakers with the lowest correlation, coming from the group of male speaker between 60 and 80 years, with 0.76 average and the female speakers in the range between 50 and 60 years with 0.57 average. This suggest that for speakers in this group further distinction would be needed.

4.2 Stagged Classification

For the two staged classification the first search was for a baseline solution. For this the whole feature set was taken and a support vector machine was trained to classify the utterances into the categories of mental underload and mental overload. As can be seen in Table 4 this began with an unweighted average recall (UAR) of 68.5% with a better internal recall for overload in comparison to underload of 71.9% and 65.1% respectively. This happens most likely because the users tend to speak more in the overload phase which leads to a greater set of overload utterances. The unweighted average precision (UAP) is 68.7% with roughly the same for each class (67.5% for underload and 68.7% for overload). A comparable experiment with RF brings results in the range of 74.2% UAR (with 86.2% and 62.2% each) and an UAP of 75.9% (with 71.7% and 80.1% each).

In the next step the amount of used features from the full set with 988 features were reduced to the set examined in the last section of 4.1. For this two strong diverging results were received, the SVM lost a great amount of precision and recall, while the RF method was relatively only slightly worse than the original results. Both results could

Table 4. The Results for Support Vector Machines (SVM) and Random Forests (RF) with all Speakers in on Group

	Full Feature Set		Reduced Feature Set	
	SVM	RF	SVM	RF
UAR	68.5	74.2	46.4	69.2
UAP	68.7	75.9	46.5	69.6

Table 5. The aggregated Results for Support Vector Machines (SVM) and Random Forests (RF) with Speakers separated into their respective sex

	Full Feature Set		Reduced Feature Set	
	SVM	RF	SVM	RF
UAR	69.5	75.5	68.4	69.8
UAP	69.2	75.6	60,2	69.9

be optimized by including further features as searched by their ranking, with the results for RFs were nearly on the original level during the top 20 features, while the results for SVM were rising slower.

For the test of separating the classification task the sex of the speakers were taken as an easy test criterion. The separation into this class was nearly perfect for RF methods, and considerably higher for SVMs than the detection of underload and overload, especially in view of the precision (RF with 97.2% UAR and 92.9% UAP and SVM with 86.4% UAR and 84.9% UAP) This takes the part of the first stage of this classifier.

For the next step the results for the classification tasks were examined with this resulting smaller sets. The results for the full feature set were found to be roughly in the same range as the former test, with both performing slightly better, as can be seen in Table 5.

The interesting development was with the reduced feature set. Nearly the same amount of recall and precision were reached compared to the original test with full set over all speakers when using the SVMs, both in range of 68% UAR while some precision was lost. This is strongly connected to the smaller set if male speakers, as the UAP for female Speakers lies at 62.1% and for male speakers at 58.8%. One can assume, that with a bigger training set, some

of these effects could be mitigated. For the RFs slightly better results were achieved than the reduced set for all speakers together, without reaching the effect of the full feature set.

5. CONCLUSION AND OUTLOOK

Several important steps were examined in this paper. First a measure of the general importance of one feature in relation to the amount of speakers in a groups was found, as this can change according to which group of examples one uses for training and testing and should be considered for future test into the personalization of classifiers for its user. Then, that certain features prevail in their distinctiveness but not in the same way for both male and female speakers. This was partially due to their different vocal range, but also in the way the speaker reacted to certain situations. Considering Question 1, there were not only changes in the range of the features but also slight changes to the most distinct features themselves. While for Question 2, separating the set in two roughly comparable groups helped for some classifications. Based on this examination, features which proved a strong correlation for most speakers were found, even though there were always subgroups for which this was not the case. This makes the use of one general set of features which is usable for all speakers unlikely. The solution of simple adding all possible features into one classifier is dependent of the inert ability of the system to distinguish the special cases, which in turn needs a complex system as given by deep neural networks or similar computational complex systems.

As shown this must not necessarily be useful when there are relevant subgroups of instances which divert from the norm the system could otherwise solve with a miniscule subset of necessary information as shown with the result for 10 or 20 chosen features instead of all 988 features. By staging the difficult decision process into several smaller ones one can potentially replace a lot of overhead otherwise necessary. In the experiments there is a distinction between the results of the SVMs and the RFs classifier. For RF, which are already working with a kind of decision tree there was no improvement on the already very good starting results, but one could see that most information was contained in the reduced feature set. With the SVMs on the other hand one could see into the advantages of this staged approach, and while the general results were not comparable to the RF method, the improvement with the reduced feature set in the staged approach and even the full feature set showed considerable promise, especially for tasks which cannot be solved efficiently by decision trees. With the reduced amount of data necessary for the classification it is easier to employ the system on mobile applications, as asked in Question 3, as instead of one big system with a lot of data capacity, one now can use smaller steps with lower data requirements.

The research has shown promising results in the ability to retain or slightly improve the results of while reducing the amount of data, this was especially strong for methods which are not inherently based on decision trees like for example SVM. For the future we will try to improve on this by either reducing the base set further, which requires bigger dataset as shown by the slightly worse results we got from male speakers from which we had less compared to female ones. Another aspect of research would be the intelligent clustering of our set, instead of the manual distinction as done in this experiment. With this we could circumvent possible overlaps which could otherwise occur and be more precise in finding the nonstandard cases in our examples. Otherwise we would like to extend this research on other database with similar structure.

ACKNOWLEDGEMENTS

We acknowledge support by the project “Intention-based Anticipatory Interactive Systems” (IAIS) funded by the European Funds for Regional Development (EFRE) and by the Federal State of Sachsen-Anhalt, Germany, under the grant number ZS/2017/10/88785.

REFERENCES

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

Böck, R., Egorow, O., and Wendemuth, A. (2018). Acoustic detection of consecutive stages of spoken interaction based on speaker-group specific features. In *Proceedings of the 29. Konferenz Elektronische Sprachsignalverarbeitung*, 247–254. Ulm, Germany.

Böck, R., Hübner, D., and Wendemuth, A. (2010). Determining Optimal Signal Features and Parameters for HMM-Based Emotion Classification. In *Proceeding of the 15th IEEE Mediterranean Eletrotechnical Conference*, 1586–1590. Valleta, Malta.

Brown, W., Morris, R.J., Hollien, H., and Howell, E. (1991). Speaking fundamental frequency characteristics as a function of age and professional singing. *Journal of Voice*, 5(4), 310 – 315.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., and Weiss, B. (2005). A database of German emotional speech. In *INTERSPEECH-2005*, 1517–1520. ISCA, Lisbon, Portugal.

de Looze, C. and Rauzy, S. (2011). Measuring speakers similarity in speech by means of prosodic cues: Methods and potential. In *INTERSPEECH-2011*, 1393–1396. ISCA, Florence, Italy.

Egorow, O. and Wendemuth, A. (2016). Detection of challenging dialogue stages using acoustic signals and biosignals. In *Proceedings of the WSCG 2016*, 137–143. ISCA, Plzen, Czech Republic.

Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., and Truong, K.P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM ’13, 835–838. ACM, New York, NY, USA.

Frank, E., Hall, M., and Witten, I. (2016). The weka workbench. online appendix for ”data mining: Practical

machine learning tools and techniques". *Morgan Kaufmann, Fourth Edition, 2016.*

Frommer, J., Michaelis, B., Rösner, D., Wendemuth, A., Friesen, R., Haase, M., Kunze, M., Andrich, R., Lange, J., Panning, A., and Siegert, I. (2012a). Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 3064–3069. European Language Resources Association (ELRA), Istanbul. Turkey.

Frommer, J., Rösner, D., Haase, M., Lange, J., Friesen, R., and Otto, M. (2012b). Dejection and Avoidance of Failures in Dialogues - Wizard of Oz Experiment Operators Manual. Pabst Science Publishers.

Griffiths, P.E. and Scarantino, A. (2009). Emotions in the wild: The situated perspective on emotion. In *The Cambridge handbook of situated cognition*, 437–453. Cambridge University Press.

Kruskal, W.H. and Wallis, W.A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260), 583–621.

Nass, C., Steuer, J., and Tauber, E.R. (1994). Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, 72–78. ACM, New York, NY, USA.

Panning, A., Siegert, I., Al-Hamadi, A., Wendemuth, A., Rösner, D., Frommer, J., Krell, G., and Michaelis, B. (2012). Multimodal affect recognition in spontaneous HCI environment. In *2012 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC 2012*, 430–435. Hong Kong, China.

Picard, R.W. (1999). Affective Computing for HCI. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I - Volume I*, 829–833. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., and Otto, M. (2012). LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 96. European Language Resources Association (ELRA), Istanbul, Turkey.

Scherer, K. (2001). Appraisal considered as a process of multilevel sequential checking. In *Appraisal Processes in Emotion: Theory, Methods, Research*, volume 92, 92–120. Oxford University Press.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. *Speech Communications*, 53(9-10), 1062–1087.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (2009). Acoustic emotion recognition: A benchmark comparison of performances. *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, 552–557.

Shahzadi, A., Ahmadyfard, A., Yaghmaie, K., and Harimi, A. (2013). Recognition of Emotion in Speech Using Spectral Patterns. volume 26, 140–158. University of Malaya.

Siegert, I., Philippou-Hübner, D., Hartmann, K., Böck, R., and Wendemuth, A. (2014). Investigation of speaker group-dependent modelling for recognition of affective states from speech. *Cognitive Computation*, 6(4), 892–913.

Siegert, I., Tang, S., and Requardt, A. (2018). Acoustic addressee-detection analysing the impact of age, gender and technical knowledge. In *Proceedings of the 29. Konferenz Elektronische Sprachsignalverarbeitung*, 113–120. Ulm, Germany.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 88–103.

Tahon, M. and Devillers, L. (2016). Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1), 16–28.

Tornow, M., Krippl, M., Bade, S., Thiers, A., Siegert, I., Handrich, S., Krüger, J., Schega, L., and Wendemuth, A. (2016). Integrated Health and Fitness (iGF)-Corpus - ten-Modal Highly Synchronized Subject-Dispositional and Emotional Human Machine Interactions. In *Proceedings Multimodal Corpora: Computer Vision and Language Procession (MMC 2016) - ELRA*, 21–24. Portorož, Spain.

Ward, R.D. and Marsden, P.H. (2004). Affective computing: Problems, reactions and intentions. In *Interacting with Computers*, volume 16, 707–713.