Active Learning for Speech Event Detection in HCI

Patrick Thiam(✉), Sascha Meudt, Friedhelm Schwenker, and Günther Palm

Institute of Neural Information Processing, University of Ulm,
James-Franck-Ring, 89081 Ulm, Germany
{patrick.thiam,sascha.meudt,friedhelm.schwenker,guenther.palm}@uni-ulm.de

Abstract. In this work, a pool-based active learning approach combining outlier detection methods with uncertainty sampling is proposed for speech event detection. Events in this case are regarded as atypical utterances (e.g. laughter, heavy breathing) occurring sporadically during a Human Computer Interaction (HCI) scenario. The proposed approach consists in using rank aggregation to select informative speech segments which have previously been ranked using different outlier detection techniques combined with an uncertainty sampling technique. The uncertainty sampling method is based on the distance to the boundary of a Support Vector Machine with Radial Basis Function kernel trained on the available annotated samples. Extensive experimental results prove the effectiveness of the proposed approach.

Keywords: Active learning · Supervised learning · Support Vector Machines · Support Vector Data Description · Gaussian Mixture Model · Rank aggregation

1 Introduction

The performance of Supervised Learning techniques relies strongly on the quality and the quantity of the available annotated data. Meanwhile, the annotation process of a huge dataset is known to be very cumbersome and expensive. This explains the scarceness of large labeled datasets, while, on the other hand, a large amount of unlabeled data is available. In the present work, we propose an active learning technique designed with the goal of attaining similar results as classifiers trained on fully annotated datasets by carefully selecting and annotating a strongly reduced subset.

Event detection can enhance the capabilities of an emotion recognition system. Once an event is detected, it can be subsequently given a label depicting the

P. Thiam—This paper is based on work done within the Transregional Collaborative Research Centre SFB/TRR 62 Companion-Technology for Cognitive Technical Systems funded by the German Research Foundation (DFG). The author is supported by the BMBF within the project *SenseEmotion*. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

F. Schwenker et al. (Eds.): ANNPR 2016, LNAI 9896, pp. 285–297, 2016.
DOI: 10.1007/978-3-319-46182-3_24

specific affective disposition of the user. For instance, laughter can be perceived as a sign of happiness or frustration. Heavy breathing can also be interpreted as a sign of despair or relief. Therefore, a cascaded classification consisting in first localizing the occurring events before proceeding with a specific classification would enhance the capabilities of the recognition system.

In this work, we focus on event detection in speech signals. Events are defined as atypical utterances in the speech signal such as laughter, heavy breathing or other expressions of dissatisfaction, frustration or despair. A pool-based active learning approach is proposed that combines uncertainty sampling with outlier detection techniques in order to select the most informative samples. By carefully selecting and annotating the samples from which a classification model is trained, the cumbersome and time expensive process of annotating the whole available dataset is avoided while no significant degradation of the performance of the generated system is observed. The proposed approach is tested on a subset of the Ulm University Multimodal Affective Corpus (uulmMAC) database, using uniquely the speech modality. We assess the effectiveness of the developed active learning approach by comparing its performance with a model trained on a fully annotated training set.

The remainder of this work is organised as follows. In the next section the utilised dataset is described, as well as the manual annotation process, followed by a description of the feature extraction process and the first assessment of the annotated set through supervised learning. The active learning approach is explained in the following section, followed by the presentation and the discussion of the experimental results. The work is finally concluded in the last section.

2 Dataset Description

The utilized dataset is a subset of nine participants of the Ulm University Multimodal Affective Corpus (uulmMAC database). The interaction scheme of uulmMAC is based on the work of Schüssel et. al. who proposed a gamified experimental setup close to everyday life Human Computer Interaction [16] (e.g. looking for a specific menu entry or button followed by selecting or interacting with this element). The very generic paradigm allows a lot of variations in order to investigate different research questions. uulmMAC consists of 60 participants in 100 recording sessions of about 45 min each. The recordings constitute of synchronous multimodal data containing three video streams (frontal HD, frontal webcam, rear webcam), three audio lines (headset, ambiance and directional microphone), MS Kinect 2 (depth, infrared, video, audio, posture) and biophysiology (EMG, ECG, SCL, respiration and temperature). Figure 1 shows an overview of the experimental setup.

In the uulmMAC participants were asked to play a series of games to test their reaction to both overwhelming and boring difficulty levels. The task of each game sequence was to identify the singleton element, i.e. the one item that is unique in shape and color (number 36 and 2 in Fig. 2). The difficulty was set by adjusting the number of shapes shown per sequence and increasingly

less money earned the longer they needed to answer. If the given answer was incorrect, the participant received no reward at all for that particular round. To modify the difficulty further, the time given per sequence was also adjusted. Each participant completed an introductory sequence and four game sequences of decreasing difficulty. The first sequence was designed to induce overload (6×6 board with 6 s to provide an answer, see Fig. 2), the second was a mix of 5×5 and 4×4 with 10 s to provide an answer, the third was set to 3×3 with 100 s to provide an answer. Sequence 4 was designed to induce underload and the game setting was therefore set to an easy 3×3 mode with 100 s answering time. The last sequence induced frustration, e.g. by purposely logging in a wrong answer.

After each accomplished sequence the participants answered a series of questions. The aim of those questions was to determine valence, arousal and dominance [15] experienced in the particular sequence. Firstly, the participants answered in their own words how they felt during the game. Afterwards each

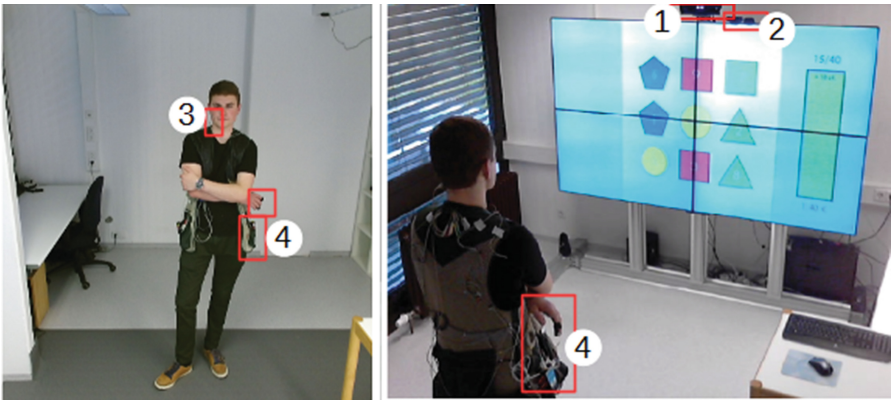

Fig. 1. The experimental setup with sensors MS Kinect 2 (1), front cameras (2), headset (3) and biosensors (4) (some physiologic sensors are placed under the shirt). **Left**: frontal camera view. **Right**: rear camera view.

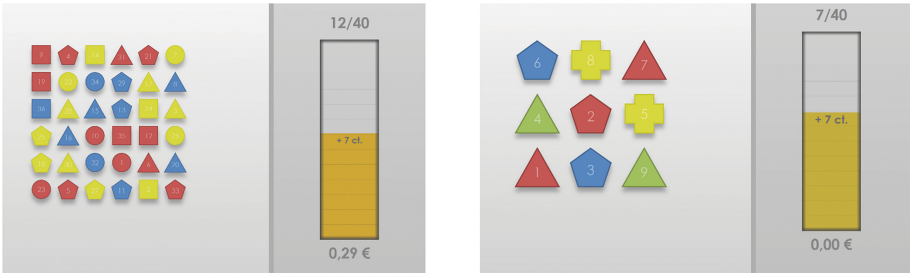

Fig. 2. The game mode. Left: a very complex game mode causing a high mental load. **Right**: a simple game mode causing the player to relax more.

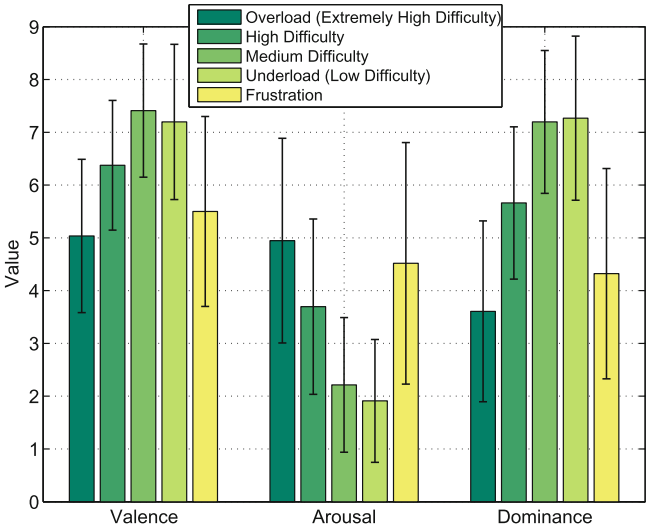

Fig. 3. Mean VAD over all game sequences. It can be seen, that the reported VAD varies significantly between the sequences, especially between sequences 1 and 4.

participant was presented with three scales and was asked to choose the value according to their experience (Self Assessment Scale (SAM)) [3].

Figure 3 shows how valence, arousal and dominance are evolving over the five sequences. The results suggest that the reported (V, A, D) experiences differ significantly between the mental overload and underload sequences. It seems that valence is higher in mental underload, arousal is higher in mental overload and dominance is higher in the mental underload sequence. According to this, mental overload and underload can be expressed using the (V, A, D) space as overload: $(V-, A+, D-)$ and underload: $(V+, A-, D+)$.

In a preliminary step of the current study, a set of 9 participants was selected based on their grade of expressiveness by observing their demeanour throughout the different phases of the experiment. The present work focuses on the headset audio channel of this reduced set of participants which consists of 4 male and 5 female participants, aged from 20 to 27 years old. Since an identical speech signal was recorded on both channels of the stereo recording from the headset, a single channel was used for further analysis.

The preliminary analysis of the speech signal showed that each candidate expressed sporadic reactions when they failed to give an accurate answer to a specific task, in particular during the phases of overload throughout which they were overwhelmed by the level of difficulty and the speed of the tasks. The present work focuses on the detection and the discrimination between such atypical reactions (e.g. laughter, heavy breathing, idiomatic expressions as a signal of boredom, dissatisfaction, frustration and despair) and the answers provided by the candidates which are principally spoken indexes of selected cells within a grid displayed during the tasks. In the present work, the atypical utterances described earlier will be referred to as events and the provided answers as normal utterances.

In order to be able to assess the performance of the developed methods, a second step was undertaken during which the extracted speech signal specific to each of the 9 candidates was manually annotated.

2.1 Manual Annotation

A manual annotation had to be carried out in order to provide ground truth labels that are essential for a further analysis and characterization of the disposition of each participant during the experiments, as well as for the proper assessment of the developed methods. The annotation step was preceded by an automatic segmentation step consisting in distinguishing between voice active segments and voice inactive (silence or noise) segments. This step was necessary since the study focuses on the aforementioned two classes of voice active segments (events and normal utterances). Since the recordings were performed in a relatively clean (noise-free) environment, a simple unsupervised voice activity detection [1] based on the energy of the speech signal was applied and yielded satisfactory results.

Subsequently, based on the results of the voice activity detection, each detected voiced segment was manually annotated as being either an event or a normal utterance. Furthermore, the boundaries of the voiced regions were manually adjusted in order to acquire a precise segmentation of the utterances

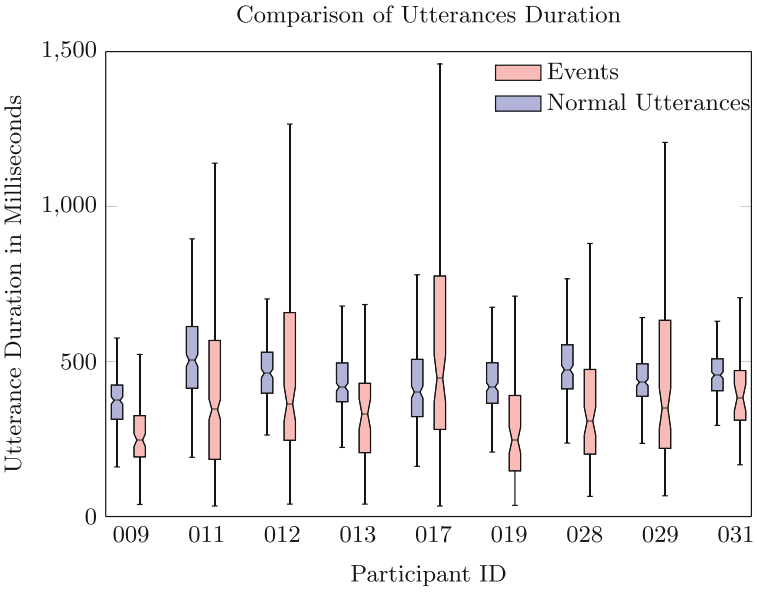

Fig. 4. Manual Annotation Results: participant dependent assessment of the duration of the annotated utterances (events and normal utterances). The figure depicts a great variation of utterance duration among the different participants.

and to substantially reduce the amount of noise in each segment. The whole annotation process was performed within the ATLAS annotation tool [14].

Following the annotation of the speech signal, a first assessment of the labeled speech samples was undertaken by comparing the duration of the normal utterances to those of events (see Fig. 4). The first observation is the specificity of the duration of the utterances to each participant. Each participant acts differently depending on his or her actual affective state. Secondly, for most of the participants the annotated events depict a greater variance in duration in comparison to normal utterances. This can be explained by the fact that the answers to the different tasks are basically pronounced numbers while events would vary from very short locutions to entire phrases spoken as a sign of frustration. Furthermore, the median duration of normal utterances is higher than the median duration of events for almost all participants. These observations describe the great diversity of the demeanour of the participants depending on their mental or affective disposition.

Moreover, the overall minimum duration of an event is situated somewhere between 30 and 50 ms, while the overall minimum duration of a normal utterance is situated somewhere between 150 and 200 ms. Based on these findings, 2 windows of respectively 115 and 215 ms were selected for the segmentation of the annotated speech samples in order to proceed with the feature extraction. The windows were shifted with a fixed offset of 65 ms. The resulting data distribution can be observed in both Tables 1 and 2. In both segmentation settings, except for one participant (Participant 011), the data distribution is strongly imbalanced, with more normal utterances (majority class) than events (minority class). Therefore, the geometric mean (gmean) [7,13] defined in Eq. 1 is used as performance metric for the assessment of the developed methods:

$$gmean = \sqrt{acc^{+} \times acc^{-}} \tag{1}$$

where acc^{+} stands for the accuracy on the minority class and acc^{-} stands for the accuracy on the majority class.

Table 1. Data Distribution: window size set at 115 ms

Participant ID	009	011	012	013	017	019	028	029	031
Events	572	2863	1174	304	1276	450	656	849	281
Normal utterances	1579	2817	2137	2062	1894	1927	2449	2082	2132
Total	2151	5680	3311	2366	3170	2377	3105	2931	2413

Table 2. Data Distribution: window size set at 215 ms

Participant ID	009	011	012	013	017	019	028	029	031
Events	350	2206	890	195	1036	297	439	632	196
Normal utterances	1033	2162	1574	1464	1330	1363	1851	1503	1555
Total	1383	4368	2464	1659	2366	1660	2290	2135	1751

2.2 Feature Extraction

Following the segmentation of the annotated speech samples, a set of audio features was extracted from the different segments with fixed frames of 25 ms length sampled at a rate of 10 ms: 8 linear predictive coding coefficients (LPC) [10]; 5 perceptual linear prediction cepstral coefficients (PLP-CC) [8], each with delta and acceleration coefficients; 12 Mel frequency cepstral coefficients (MFCC) [9], each with delta and acceleration coefficients; fundamental frequency (F0); voicing probability; loudness contour; log-energy with its delta and acceleration [11]. Thus, each frame is represented by a 65 dimensional feature vector. Consequently, each labeled speech segment is represented by a 10×65 feature matrix for a window size of 115 ms, and 20×65 feature matrix for a window size of 215 ms. The features were extracted using the openSMILE feature extraction tool [6].

Subsequently, in order to compute the features at the segment level, 14 statistical functions (mean, median, standard deviation, maximum, minimum, range, skewness, kurtosis, first and second quartile, inter quartile, 1 %-percentile and 99 %-percentile, range of 1 % and 99 % percentile), were applied on each of the extracted frame level feature resulting in a 910 dimensional feature vector for each segment. These segment level feature vectors are used for the assessment of the developed methods.

2.3 Dataset Assessment: Supervised Learning Experiments

An important step of the study is the assessment of the generated dataset in order to ensure the relevance of the extracted feature vectors for the task at hand. Therefore, 3 classifiers were selected to perform a 5-fold participant dependent blocked cross validation [2] of the dataset for each segmentation setting: a Random Forest classifier with a fixed size of 300 trees; a Support Vector Machine (SVM) with a Gaussian Radial Basis Function (RBF) kernel; a multilayer perceptron (MLP) with a unique hidden layer of 100 neurons, each with a radial basis transfer function.

During the blocked cross validation the dataset is partitioned sequentially (not randomly) into several subsets. Therewith overlapping segments belonging to a specific utterance are all contained either in the training or in the testing set except for a few segments situated at the border of the partitions. Since we perform a 5-fold cross validation, such occurrences are minimal. Thus they are not discarded and kept in the sets. Following, each subset is used within each cross validation iteration as a test set while the remaining sets are used as training sets. Before the classification, each set is first preprocessed to get rid of noisy data. Subsequently, Principal Component Analysis (PCA) is performed to reduce the dimensionality of the dataset. The first 10 components are used at each iteration. Then the Synthetic Minority Over-Sampling Technique (SMOTE) [5] is applied on the training set to balance the data before the learning process takes place.

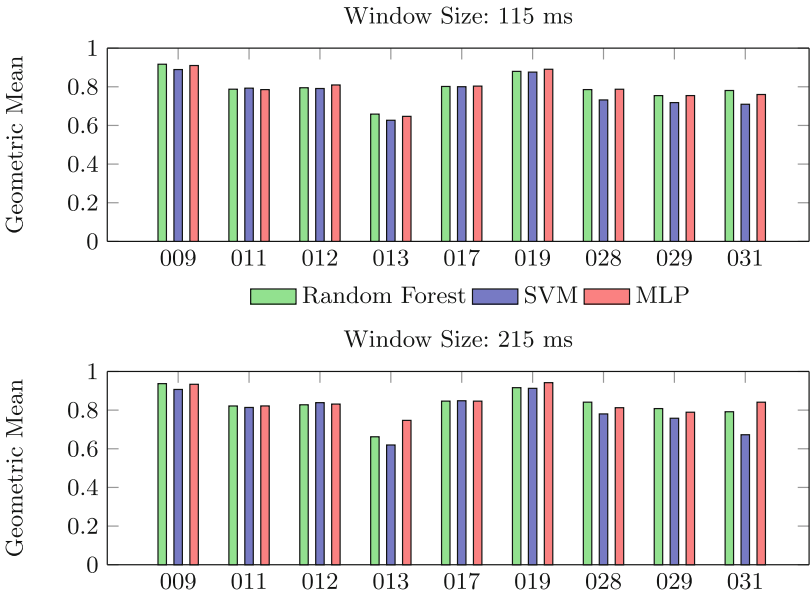

Fig. 5. Supervised Learning Results: person dependent 5-fold blocked cross validation. The upper figure depicts the results for a window size of 115 ms, while the lower figure depicts the results for a window size of 215 ms. Both figures depict similar performances amongst the participants despite the different window sizes with the Random Forest classifier performing in most cases slightly better than the MLP and the SVM.

The classification results can be seen in Fig. 5. The upper figure shows the results for the window size of 115 ms. The lower figure shows the results for the window size of 215 ms. The selected classifiers perform rather similarly in both segmentation settings with the best performances achieved by the Random Forest classifier in most cases, followed by the MLP and finally the SVM. These results validate the relevance of the generated feature set, as well as the applicability of different classifiers, to this specific speech event detection task. Based on these findings, an active learning method is proposed with the goal of reaching similar classification performances as depicted in Fig. 5, while reducing the costs of annotating the whole set.

3 Proposed Active Learning Approach

We propose a pool-based active learning approach that combines outlier detection with uncertainty sampling to select the most informative samples that are annotated by an oracle and subsequently used to train a classification model. The proposed approach is a further iteration of the methods presented by Thiam et al. in [18,19].

More specifically, the method consists in first selecting and ranking a subset of samples from the unlabeled set based on outlier detection techniques. Secondly, a SVM with a RBF kernel is trained on the available labeled set and used to select and rank another subset of samples from the same unlabeled set based on the distance from those samples to the decision boundary of the trained model. Subsequently, rank aggregation is performed on the resulting subsets and the k highest ranked samples are selected for the annotation process. Two outlier detection techniques have been experimented with.

3.1 Outlier Detection with an Ensemble of Support Vector Data Description

This approach consists in generating an ensemble of Support Vector Data Description (SVDD) [17] models and performing the selection and ranking of the samples based on the voting count of the ensemble. As opposed to the method presented by Thiam et. al. [18], the models of the ensemble are not randomly generated. Based on the work of Chang et al. [4], a total of $m \times n$ SVDD models are generated by choosing m values for the parameter C, equally spaced within the interval $\left[\frac{1}{N}, 1\right]$ where N is the number of samples in the unlabeled set, and n values for the parameter γ of the RBF kernel, equally spaced within the interval $\left[\frac{1}{f}, 1\right]$ where f is the dimensionality of the feature vectors. In this way, the grid of possible values for both parameters is covered and the diversity in the ensemble is ensured. Furthermore a threshold is used to prune the generated ensemble, based on the reclassification results of each generated model. Models with outlier classification rates higher than the specified threshold are discarded. Subsequently, the samples of the unlabeled set are ranked in descending order of the voting count.

3.2 Outlier Detection with a Gaussian Mixture Model

This approach consists in generating a Gaussian Mixture Model (GMM) from the unlabeled set and using the Mahalanobis distance from each sample to each Gaussian component center to determine the outliers. A specified threshold is also used in this case for the purpose of detecting outliers. Samples having a Mahalanobis distance higher than the specified threshold are considered as outliers. Once this has been done for each component, each sample in the subset of detected outliers is ranked in ascending order of the probability density function of the GMM.

3.3 Rank Aggregation and Sample Selection

The selection of the samples to be annotated is performed by applying Borda's geometric-mean rank aggregation method [12] on both subsets resulting from the outlier detection technique and the SVM model trained on the available labeled set and applied on the unlabeled set. Subsequently, the k highest ranked samples are selected for the annotation.

3.4 Experimental Settings and Results

The proposed approach is tested in a 5-fold blocked cross validation setting during which after the sequential partition of the dataset is performed, each partition is used within each cross validation iteration as a test set and the remaining partitions are used as training set. During each iteration, the baseline is first computed by training a SVM with RBF kernel on the training set and performing the classification using the generated model on the test set. Next, the proposed approach is applied on the training set. After each active learning iteration, a SVM with RBF kernel is trained on the annotated samples and applied on the test set. A fixed number of 25 SVDD models ($m = 5, n = 5$) is generated for the committee and the number of components for the GMM is fixed at 5. Moreover, a maximum of 50 samples is selected during each active learning round.

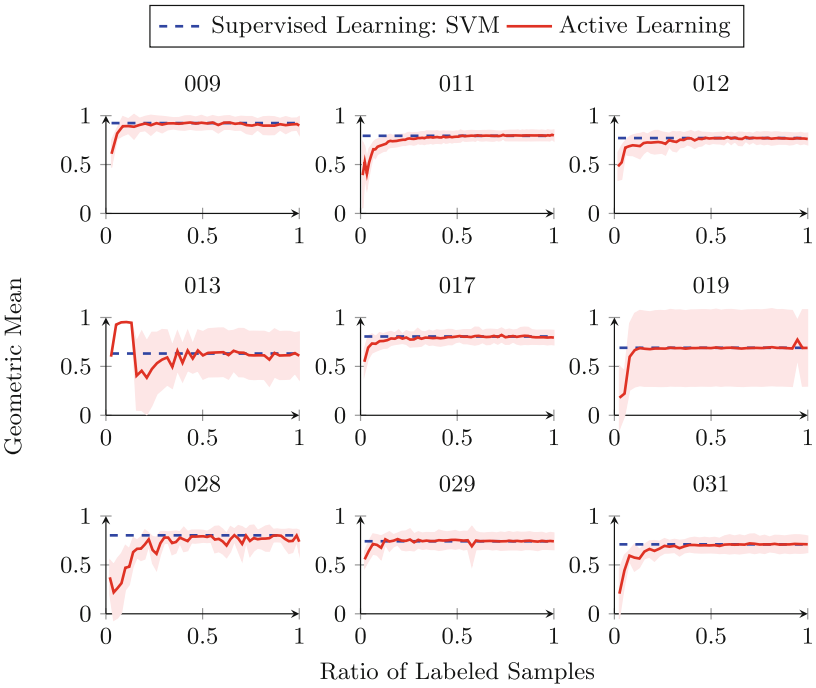

Fig. 6. Active Learning with SVDD Outlier Detection Results: person dependent 5-fold blocked cross validation with a window size of 115 ms. The dashed blue line corresponds to the averaged geometric mean of the fully supervised learning results computed with a SVM with RBF kernel. The continuous red line corresponds to the averaged supervised learning results computed with a SVM with RBF kernel trained on the available annotated samples after each active learning round. The lighter coloured corridors correspond to the standard deviation of the results after each active learning round. (Color figure online)

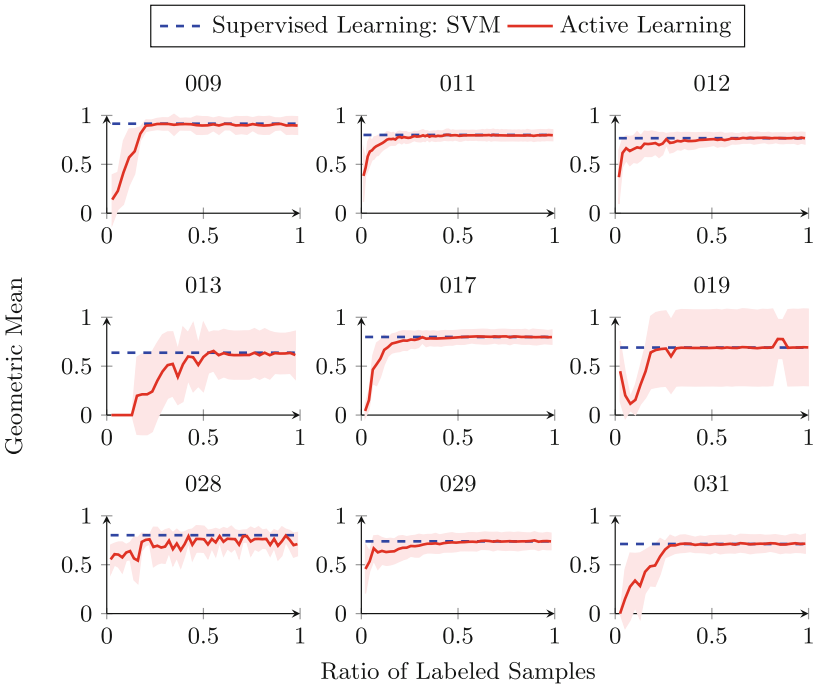

Fig. 7. Active Learning with GMM Outlier Detection Results: person dependent 5-fold blocked cross validation with a window size of 115 ms. The dashed blue line corresponds to the averaged geometric mean of the fully supervised learning results computed with a SVM with RBF kernel. The continuous red line corresponds to the averaged supervised learning results computed with a SVM with RBF kernel trained on the available annotated samples after each active learning round. The lighter coloured corridors correspond to the standard deviation of the results after each active learning round. (Color figure online)

Figures 6 and 7 show the results of the approach, based respectively on the ensemble of SVDD models and the GMM, applied on the data specific to each participant, with the window size of 115 ms (the same experiments have been conducted with the window size of 215 ms yielding similar results). The averaged geometric mean of the supervised learning classification of the blocked cross validation is plotted in blue and serves as the baseline in each case. The averaged geometric mean of the supervised learning classification performed using the annotated samples after each active learning round and applied on the test set is plotted in red. Furthermore, the standard deviation of the classification is plotted in lighter coloured corridors to depict the variance of the achieved results after each active learning round.

The figures specific to the participants 013 and 019 depict a great degree of variance during the whole active learning process in both outlier detection settings. This is due to the highly imbalanced distribution of events within each

partition generated by the sequential cross validation, causing a high fluctuation of the chosen performance metric (geometric mean). For those participants where the distribution of events in each sequence is more balanced, the variance gradually sinks with each round of active learning until a certain point, then remains constant until the final iteration. This observation shows that the system remains stable once the most interesting samples are annotated.

The figures also depict the performance of both outlier detection methods. In both cases and for most of the participants, the baseline is already attained with less than half of the whole training set being annotated, proving the effectiveness of the proposed approach. A thorough comparison of both outlier detection methods has not been undertaken in the present work.

4 Conclusion and Future Work

In this work, we proposed an active learning approach for the annotation and detection of events in speech signals. The approach consists in combining outlier detection methods and uncertainty sampling based on the decision boundary of a SVM with RBF kernel to select the most informative samples to be annotated. The method has been assessed on a subset of the uulmMAC dataset using solely the speech modality and proved to be effective, since for most of the assessed participants, a little less than half of the training set had to be annotated to attain the same classification performance as a model trained on the fully annotated training set.

An application of the designed method in HCI could be the annotation and classification of a specific category of emotion (e.g. anger, sadness, happiness) in a one against all scenario, without having to go through the process of annotating an entire dataset. Another application of the method would be an exploratory and cascaded annotation of a dataset consisting in first detecting various atypical states in the dataset before providing specific labels to the detected events in a further step.

Further experiments are to be undertaken, involving the assessment of the applicability of the presented approach on other modalities (e.g. video and biophysiological signals). Moreover, several outlier detection methods are also to be assessed, as well as different combination methods for the selection of the most informative samples.

References

1. Alam, M.J., Kenny, P., Ouellet, P., Stafylakis, T., Dumouchel, P.: Supervised/unsupervised voice activity detector for text-dependent speaker recognition on RSR2015 corpus. In: Odyssey Speaker and Language Recognition Workshop (2014)
2. Bergmeir, C., Benìtez, J.M.: On the use of cross-validation for time series predictor evaluation. Inf. Sci. **191**, 192–213 (2012)
3. Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994)

4. Chang, W.C., Lee, C.P., Lin, C.J.: A revisit to support vector data description (SVDD). Technical reports (2013)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
6. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in opensmile, the Munich open-source multimedia feature extractor. In: ACM Multimedia (MM), pp. 835–838, October 2013
7. Gu, Q., Zhu, L., Cai, Z.: Evaluation measures of the classification performance of imbalanced data sets. In: Cai, Z., Li, Z., Kang, Z., Liu, Y. (eds.) ISICA 2009. CCIS, vol. 51, pp. 461–471. Springer, Heidelberg (2009)
8. Hermansky, H.: Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Am. **87**(4), 1738–1752 (1990)
9. Jagan Mohan, B., Ramesh Babu, N.: Speech recognition using MFCC and DTW. In: 2014 International Conference on Advances in Electrical Engineering (ICAEE), pp. 1–4, January 2014
10. Krothapalli, S.R., Koolagudi, S.G.: Emotion recognition using vocal tract information. In: Krothapalli, S.R., Koolagudi, S.G. (eds.) Emotion Recognition using Speech Features. SpringerBriefs in Electrical and Computer Engineering, pp. 67–78. Springer, New York (2013)
11. Krothapalli, S.R., Koolagudi, S.G.: Speech emotion recognition: a review. In: Krothapalli, S.R., Koolagudi, S.G. (eds.) Emotion Recognition using Speech Features. SpringerBriefs in Electrical and Computer Engineering, pp. 15–34. Springer, New York (2013)
12. Lin, S.: Rank aggregation methods. Wiley Interdisc. Rev. Comput. Stat. **2**(5), 555–570 (2010)
13. Lòpez, V., Fernàndez, A., Garcìa, S., Palade, V., Herrera, F.: Strategies for learning in class imbalance problems. Pattern Recogn. **36**(3), 849–851 (2003)
14. Meudt, S., Bigalke, L., Schwenker, F.: Atlas - an annotation tool for HCI data utilizing machine learning methods. In: Proceedings of the 1st International Conference on Affective and Pleasurable Design (APD 2012) (Jointly with the 4th International Conference on Applied Human Factors and Ergonomics (AHFE 2012)), pp. 5347–5352 (2012)
15. Russel, J.A.: Core affect and the psychological construction of emotion. Pyschological Rev. **110**(1), 145–172 (2003)
16. Schüssel, F., Honold, F., Bubalo, N., Huckauf, A., Traue, H., Hazer-Rau, D.: In-depth analysis of multimodal interaction: an explorative paradigm. In: Kurosu, M. (ed.) HCI 2016. LNCS, vol. 9732, pp. 233–240. Springer, Heidelberg (2016)
17. Tax, D.M., Duin, R.P.: Support vector data description. Mach. Learn. **54**(1), 45–66 (2004)
18. Thiam, P., Kächele, M., Schwenker, F., Palm, G.: Ensembles of support vector data description for active learning based annotation of affective corpora. In: 2015 IEEE Symposium Series on Computational Intelligence, pp. 1801–1807, December 2015
19. Thiam, P., Meudt, S., Kächele, M., Palm, G., Schwenker, F.: Detection of emotional events utilizing support vector methods in an active learning HCI scenario. In: Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems, ERM4HCI 2014, pp. 31–36. ACM, New York (2014)