Computer Speech & Language 65 (2021) 101119

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

# Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods

Fasih Haider*[,a], Senja Pollak[b], Pierre Albert[a], Saturnino Luz[a]

[a] Usher Institute, Edinburgh Medical School, the University of Edinburgh, Edinburgh, UK
[b] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

ARTICLE INFO

*Article History:*
Received 1 February 2019
Revised 26 May 2020
Accepted 28 May 2020
Available online 1 June 2020

*Keywords:*
Feature engineering
Feature selection
Emotion recognition
Affective computing
Prosodic analysis
Cognitive health monitoring,

ABSTRACT

Research in automatic affect recognition has seldom addressed the issue of computational resource utilization. With the advent of ambient intelligence technology which employs a variety of low-power, resource-constrained devices, this issue is increasingly gaining interest. This is especially the case in the context of health and elderly care technologies, where interventions may rely on monitoring of emotional status to provide support or alert carers as appropriate. This paper focuses on emotion recognition from speech data, in settings where it is desirable to minimize memory and computational requirements. Reducing the number of features for inductive inference is a route towards this goal. In this study, we evaluate three different state-of-the-art feature selection methods: Infinite Latent Feature Selection (ILFS), ReliefF and Fisher (generalized Fisher score), and compare them to our recently proposed feature selection method named 'Active Feature Selection' (AFS). The evaluation is performed on three emotion recognition data sets (EmoDB, SAVEE and EMOVO) using two standard acoustic paralinguistic feature sets (i.e. eGeMAPs and emobase). The results show that similar or better accuracy can be achieved using subsets of features substantially smaller than the entire feature set. A machine learning model trained on a smaller feature set will reduce the memory and computational resources of an emotion recognition system which can result in lowering the barriers for use of health monitoring technology.



## 1. Introduction

Speech signals are used in a number of automatic prediction tasks, including cognitive state detection (Akira et al., 2015), cognitive load estimation (Schuller et al., 2014), presentation quality assessment (Haider et al., 2016a) and emotion recognition (El Ayadi et al., 2011; Schuller et al., 2011). Emotional/affective states could have influence on health and intervention outcomes. Positive emotions have been linked with health improvement, while negative emotions may have negative impact (Consedine and Moskowitz, 2007). For example, long term bouts of negative emotions are predisposing factors for depression (ibid.), while positive emotions-related humour and optimism have been linked with positive effects on the immune system and cardiovascular health (Dimsdale, 2008). Emotion recognition has been used in applications in the domain of health technologies, including mental health assessment and beyond (Valstar et al., 2013; Desmet and Hoste, 2013; Haider et al., 2020).

Applications using speech usually extract emotions as an additional signal in complex systems, such as in ambient intelligence (AmI) (Mano et al., 2016), depression recognition (Desmet and Hoste, 2013), and longitudinal cognitive status assessment (Su and Luz, 2016). These approaches employ very high-dimensional feature spaces consisting of large numbers of potentially relevant

*Corresponding author.
*E-mail address:* fasih.haider@ed.ac.uk (F. Haider).

https://doi.org/10.1016/j.csl.2020.101119
0885-2308/

acoustic features, usually obtained by applying statistical functionals to basic, energy, spectral and voicing related acoustic descriptors (Eyben et al., 2009) extracted from speech intervals lasting a few seconds (Ververidis and Kotropoulos, 2006). Although there is no general consensus on what the ideal set of features should be, this "brute-force" approach of employing as many features as possible seems to outperform alternative (Markovian) approaches to modelling temporal dynamics on the classifier level (Weninger et al., 2013). However, the use of such high-dimensional data sets poses challenges for prediction, as they suffer from the so-called "curse of dimensionality", high degree of redundancy in the feature set, and a large number of features with poor descriptive value. Su and Luz, for instance, noted that in a cognitive load prediction data set about 4% of a feature set of over 250 features had a standard deviation of less than 0.01 and therefore contributed negligibly to the classification task (Su and Luz, 2016). Moreover, processing of very large numbers of features presents computational challenges for the low-power, low-cost devices such as the Raspberry Pi Zero, [1] which are often used in AmI applications.

The main contribution of this study is the evaluation of different state of the art feature selection methods, including our Active Feature Selection (AFS) method, on the emotion recognition from speech, which has, to the best of our knowledge, not yet been systematically explored. This study extends our previous work (Haider et al., 2018b), where we first introduced the novel AFS method and tested it on the ICMI Challenge on Eating Conditions Recognition (Schuller et al., 2015).

## 2. Background and related work

The automatic identification of emotions in speech is a challenging task, and identifying relevant acoustic features and systematic comparative evaluations has proved difficult (Anagnostopoulos et al., 2015). In 2016, the eGeMAPs set (Eyben et al., 2016) (see Section 4.2) was designed based on features' potential to reflect affective processes and their theoretical significance. It was proposed to set a common ground of emotion-related speech features, which has become since then a *de-facto* standard. The set of target emotions has mostly been fixed around the 'Big Six', and similarly, evaluations are more and more frequently performed on a number of publicly available corpora (see Section 4.1). In the health domain, feature selection methods for speech processing have been applied to determine the most discriminant features in support of automatization efforts, as for instance in the assessment of patients with pre-dementia and Alzheimer's disease (Knig et al., 2015; Haider et al., 2020) or for the detection of sleep apnea (Goldshtein et al., 2011). The automatic emotion recognition task has gained attention in the past few years (Dhall et al., 2018; 2017). This task has been addressed through processing of facial, speech, body movements and biometric information (Knyazev et al., 2017; Haider et al., 2016b; Madzlan et al., 2015; Hu et al., 2017; Akira et al., 2015). Numerous studies (Knyazev et al., 2017; Haider et al., 2016b; Hu et al., 2017; Vielzeuf et al., 2017; Wang et al., 2017; Ouyang et al., 2017) extract audio features with OpenSMILE using *de-facto* standard presets: IS10, GeMAPS, eGeMAPs, Emobase.

The reviewed literature suggests that although the accuracy of various machine learning approaches in this area is promising, automatic dimensionality reduction has focused largely on the removal of noisy or redundant features, with less attention paid to computational resource utilisation (Akira et al., 2015; Knyazev et al., 2017; Haider et al., 2016b; Madzlan et al., 2015; Hu et al., 2017; Vielzeuf et al., 2017; Wang et al., 2017; Ouyang et al., 2017).

There are many dimensionality reduction methods: some are feature selection methods which require labelled data, and some are feature transformation methods which do not require labelled data. The former includes methods such as correlation based feature selection and Fisher feature selection (Hall, 1999; Gu et al., 2012), while the latter includes, for instance, principal component analysis (PCA), independent component analysis (ICA) (Wang and Chang, 2006) among others. Recently, efforts have focused on reducing dimensionality using PCA to improve the results for emotion recognition from speech (Jagini and Rao, 2017; Aher et al., 2016; Wang et al., 2010; Haider et al., 2018a) in different settings such as noisy setting (Aher et al., 2016). Dimensionality reduction using feature selection methods, on the other hand, are less explored in this area.

## 3. Feature selection methods

In this section we will briefly describe the feature selection methods used in this study along with our AFS method. We have selected three state of the art feature selection methods. The motivation behind using these methods here is their robust performance in a number of tasks (Roffo et al., 2017).

### 3.1. Infinite latent feature selection (ILFS)

The ILFS method (Roffo et al., 2017) performs cross-validation on an unsupervised ranking of features. At a pre-processing stage, each feature is represented by a descriptor reflecting how discriminative it is. A probabilistic latent graph containing each feature is built. Weighted edges model pairwise relations among feature distributions, created using probabilistic latent semantic analysis. The relevance of each feature is computed by looking on its weight in arbitrary set of cues. Each path in the graph represents a selection of features. The final ranking of each feature looks at its redundancy in all the possible feature subsets, selecting the most discriminative and relevant features. The evaluation on a range of different tasks (e.g. object recognition classification and DNA microarray analysis) confirms its robustness, outperforming other methods on robustness and ranking quality (Roffo et al., 2017).

[1] https://www.raspberrypi.org/products/raspberry-pi-zero/ (last accessed January 2019)

### 3.2. ReliefF

The ReliefF algorithm (Kononenko et al., 1997) which is an adaptation of the Relief feature selection method (Kira et al., 1992), performs ranking and selection of top scoring features based on their processed score. The score is calculated by weighting features on a random sample of instances. For each instance, the weight vector represents the relevance of each feature amongst the class labels: neighbours are selected from the same class (nearest hits) and from each different class (nearest misses). The weight of each feature increases when the difference with its nearest hits is low and with its nearest misses is high. Each weight vector is combined in a global relevance vector. The final subset is constituted of all the features with relevance greater than a manually set threshold. ReliefF is a common method of Feature Selection which has been continuously improved since its first publication (Kira et al., 1992; Robnik-Šikonja and Kononenko, 1997).

### 3.3. Generalized Fisher score (Fisher)

The generalized Fisher score (Gu et al., 2012) is a generalization of the Fisher score to take into account redundancy and combination of features. A subset of features is sought which maximizes the lower bound of the traditional Fisher score. A combination of features is evaluated, and redundant features discarded. A quadratically constrained linear programming (QCLP) is solved with a cutting plane algorithm. At each iteration, a multiple kernel learning is solved by a multivariate ridge regression followed by a projected gradient descent to update the kernel weights. The method produces state of the art results, outperforming many feature selection methods while having a lower complexity (Gu et al., 2012).

### 3.4. Active feature selection method

An Active Feature Selection method, which divides a feature set into subsets, has been recently introduced (Haider et al., 2018b). The term 'active' is used because compared to other approaches it evaluates feature subsets and not each feature separately, so that different features actively contribute to the feature selection. While clustering is employed, AFS does not cluster instances but dimensions. Our hypothesis is that noisy features have common characteristics that differ from those of informative features, and that clustering will divide the features into subsets according to such common characteristics. This involves clustering the data set into $N$ clusters (where $N = 5, 10, 15, \ldots, 100$) using self-organizing maps (SOM) with 200 iterations and batch training (Kohonen, 1998), and then evaluating the discrimination power of the features from each cluster $C_N$ using leave one subject out (LOSO) cross-validation, as shown in Fig. 1. The cluster with the highest validation accuracy is selected (see Fig. 6 in Section 5).

## 4. Experimentation

The section describes the datasets and their characteristics along with acoustic feature extraction and classification methods.

### 4.1. Data sets

Three corpora were selected for their shared characteristics and public availability: EmoDB, SAVEE, and EMOVO. They consist of recorded acted performances, annotated using the well-known and widely used *Big Six* set of annotations : anger, disgust, fear, happiness, sadness, surprise + neutral, except in the older EmoDB data set where boredom was used instead of surprise. Their characteristics are summarised in Tables 1 and 2.

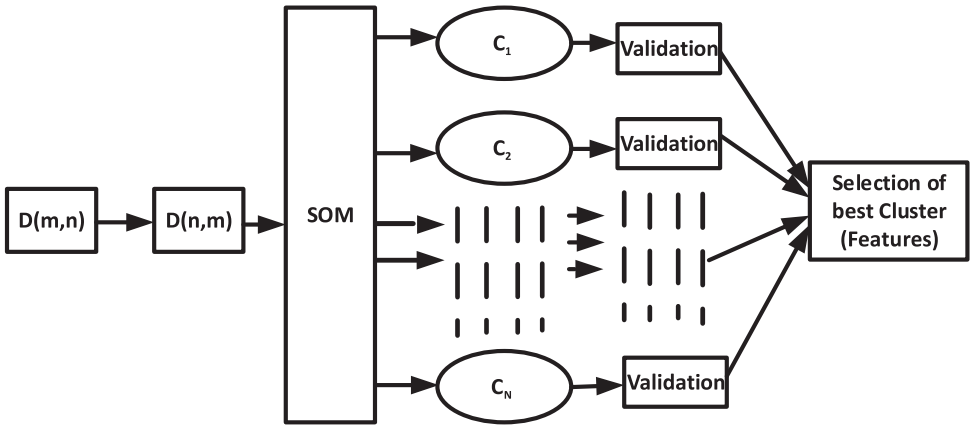


**Fig. 1.** Active feature selection method: D(m,n) represents the data where $m$ is the total number of training instances and $n$ is the total number of dimensions (988 for *emobase* and 88 for *eGeMAPs*) (Haider et al., 2018b).

**Table 1**
Main characteristics of the data sets.

| Corpus | Size (utterances) | Population | Participants | Emotion categories |
|---|---|---|---|---|
| EmoDB | 535 | 10 (5 males, 5 females) | German native speakers actors | anger, disgust, fear, joy, sadness, *boredom* + neutral |
| SAVEE | 480 | 4 (males) | English native speakers actors | anger, disgust, fear, happiness, sadness, *surprise* + neutral |
| EMOVO | 588 | 6 (3 males, 3 females) | Italian native speakers actors | anger, disgust, fear, happiness, sadness, *surprise* + neutral |

**Table 2**
Distribution of recordings across emotion categories.

| Corpus | Neutral | Anger | Disgust | Fear | Happiness | Sadness | Surprise | Boredom |
|---|---|---|---|---|---|---|---|---|
| EmoDB | 79 | 127 | 46 | 69 | 71 | 62 | - | 81 |
| SAVEE | 120 | 60 | 60 | 60 | 60 | 60 | 60 | - |
| EMOVO | 84 | 84 | 84 | 84 | 84 | 84 | 84 | - |

*Berlin Database of Emotional Speech (EmoDB)*

The EmoDB corpus (Burkhardt et al., 2005) is a data set commonly used in the automatic emotion recognition literature. It features 535 acted emotions in German, based on utterances carrying no emotional bias. The corpus was recorded in a controlled environment resulting in high quality recordings. Actors were allowed to move freely around the microphones, which affected absolute signal intensity. In addition to the emotion, each recording was labelled with phonetic transcription using the SAMPA phonetic alphabet, emotional characteristics of voice, segmentation of the syllables, and stress. The quality of the data set was evaluated by perception tests carried out by 20 human participants. In a first recognition test, subjects listened to a recording once before assigning one of the available categories, achieving an average recognition rate of 86%. A second naturalness test was performed. Documents achieving a recognition rate lower than 80% or a naturalness rate lower than 60% were discarded from the main corpus, reducing the corpus to 535 recordings from the original 800.

*Surrey Audio-Visual Expressed Emotion (SAVEE)*

SAVEE (Haq and Jackson, 2009) is an audio-visual data set that was recorded to support the development of an automatic emotion recognition system. The corpus is a set of 480 British English utterances. Each actor was recorded for 15 utterances per emotion (3 common utterances recorded for each of the 7 emotions, 2 emotion specific, and 10 generic sentences different for each emotion) and 30 neutral recordings (the 3 common and every emotion specific sentences). No limitation regarding audio features (e.g. absolute signal intensity) is explicitly stated in the description of the data set. A qualitative evaluation of the database was run as a perception tests by 10 human subjects. The mean classification accuracy for the audio modality was 66.5%, 88% for the visual modality, and 91.8% for the combined audio-visual modalities.

*Italian Emotional Speech Database (EMOVO)*

The EMOVO corpus (Costantini et al., 2014) is a speech data set featuring recorded emotions from acted performances by 6 persons. Actors were allowed to move freely around the microphones and the volume was manually adjusted, affecting absolute signal intensity. A qualitative evaluation was performed using a discrimination test. Two phrases were selected and, for each, 12 subjects had to choose between two proposed emotions. The mean accuracy for the test was about 80%.

### 4.2. Volume normalization and feature extraction

We have normalized all the speech utterances' volume into the range [-1:+1] dBFS before any acoustic feature extraction. The motivation for this is to improve the model's robustness against different recording conditions such as distance between microphone and subject. We use the openSMILE (Eyben et al., 2013) toolkit for the extraction of two acoustic feature sets which are widely used for emotion recognition. These are:

**emobase:** this acoustic feature set contains the MFCC, voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features along with their first and second order derivatives. In addition, many statistical functions are applied to these features, resulting in a total of 988 features for every speech utterance.

**eGeMAPs:** this feature set contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, hammarberg index and slope V0 features including many statistical functions applied to these features, which result in a total of 88 features for every speech utterance (Eyben et al., 2016).

### 4.3. Classification method

Classification is performed using Support Vector Machines (SVM) with a linear kernel, SMO solver and cost parameter (box constraint) set to 0.75. This classifier is employed in MATLAB[2] using the statistics and machine learning toolbox. The feature selection methods are evaluated through LOSO cross-validation, and unweighted average recall (UAR) results are computed.

[2] http://uk.mathworks.com/products/matlab/ (Last accessed: January 2019)

### 4.4. Evaluation criterion

All of the emotion recognition data sets are labeled for seven classes and we have evaluated the classifier using UAR, which corresponds to the average accuracy of all classes. The UAR measure is selected because the datasets are not balanced for emotions. The method with the highest UAR is considered the best. The blind/majority guess for this task results in a 14.3% UAR. As our focus is on feature selection methods, we set the baseline as UAR obtained using the entire feature set.

## 5. Results and discussion

We have evaluated the three different automatic feature selection methods (ILFS, ReliefF and Fisher) along with our AFS method using two different acoustic feature sets extracted from three different data sets. The results of three feature selection methods are shown in Fig. 2. The AFS results are not plotted there but in Fig. 7, as the AFS does not operate on features iteratively, but on subsets of features determined through SOM. It can be observed that around 30 out of 88 eGeMAPs features and around 100 out of 988 emobase features are sufficient to provide almost the same UAR as the highest achieved UAR for the three data sets. The best results of each feature selection method are shown in Table 3.

The results confirm that a higher accuracy can be achieved using a subset of the feature set than when using the full feature set. The results for each data set can be summarised as follows:

1. EmoDB: the ILFS method provides better UAR (69.7% for eGeMAPs and 76.9% for emobase) results than the other methods and is able to reduce the number of features (74 out of 88 for eGeMAPs and 685 out of 988 for emobase). The confusion matrix of the best UAR (76.9%) is shown in Fig. 3. For eGeMAPs, the AFS method provides an UAR of 68.5% (around 1% lower than ILFS) using 81 features. For emobase, AFS method provides an UAR of 75.8% (around 1% lower than ILFS) using 696 features. With a subset of the eGeMAPs feature set, the reliefF and Fisher methods are not able to improve over the baseline in terms of UAR. However, Fig. 2 shows that reliefF and Fisher achieved almost the same UAR as compared to baseline with only 35 eGeMAPs features instead of 88 eGeMAPs features. Hence around 60% reduction in number of features is observed.
2. EMOVO: the Fisher method yields the best UAR (41.0%) using only 25 out of 88 eGeMAPs features, while ReliefF method yields the best UAR (37.1%) for emobase (selecting 348 out of 988 features). The confusion matrix of the best UAR (41.0%) is shown in Fig. 4. The results for AFS are slightly lower than the best method (around 2%), but the number of features are significantly lower, compared to other methods. AFS selects only 2 eGeMAPs features out of 88, and 56 emobase feature out of 988, while still reaching an UAR of 39.0% and and 36.4%, respectively.
3. SAVEE: the Fisher method again yields the best UAR for eGeMAPs (34 features, and UAR of 42.4%) and emobase (158 features and UAR of 42.4%).
   The confusion matrix of the best result (UAR = 42.4%) using eGeMAPs features is shown in Fig. 5. For eGeMAPs, the results of AFS are slightly lower than the best method (around 2%). For emobase, AFS method yields and UAR of 37.5% (around 5% lower than Fisher) using 21 features.

The machine learning models trained using EmoDB (UAR=76.9%) data provide better UAR than EMOVO (41.0%) and SAVEE (42.9%). This could be due to very high quality nature of the EmoDB data set. The EmoDB data set quality was evaluated by 20 human coders with an average recognition rate of 86%, and audio recordings with the inter-coder agreement below 80% were removed (no such measure was taken for EMOVO and SAVEE).

For EMOVO, while the reported accuracy for the test set is 80% (see Section 4.1), one should note that rather than evaluating the full EMOVO data set only two phrases were selected and each coder had to choose only between two proposed emotions rather than seven. The fact that our machine learning approach to EMOVO classification is of a seven-class problem explains the much lower results obtained in comparison to human perfomance.

For SAVEE, 10 human subjects evaluated the data set and came up with an accuracy of 66.5% for audio. Our machine learning based models provide promising results as compared to humans subjects. Although they are less accurate than human annotators, we use only acoustic information to automate the process of emotion recognition, while human annotators used both acoustic and linguistic information (i.e. the spoken content).

As shown in Table 3, Generalized Fisher score provides better results in 3 out of 6 cases, ILFS provides better results in 2 out of 6 cases and reliefF provides better results in 1 out of 6 cases, indicating that overall Fisher feature selection provides the best results for the emotion recognition task.

The AFS method comes second in 3 out of 6 cases as shown in Table 3. It is also observed that the AFS method provides almost the same results in terms of UAR as the other state of the art feature selection methods, with smaller numbers of dimensions on average. We have note that for the SAVEE data set only 2 out of 88 eGeMAPs features (selected by AFS) provide better results than reliefF, ILFS and the baseline (i.e. entire feature set). For further insight into these results, we show the evaluation of clusters (feature subsets) using AFS in Fig. 6. In this figure we see that there are many clusters which provide better results than the blind guess (14.3%), while the feature cluster selected by AFS contains only 2 features (*hammarbergIndexV_sma3nz_amean* and *hammarbergIndexV_sma3nz_amean*) and leads to the 39.0% UAR. One of the possible lines of future work is to combine features from different clusters to see if this leads to improvement in classification. The AFS method was also evaluated with different numbers of clusters as shown in Fig. 7. The best UAR is obtained using 70 clusters for EMOVO dataset. The UAR values for these 70 clusters with their respective numbers of features are shown in Fig. 6.

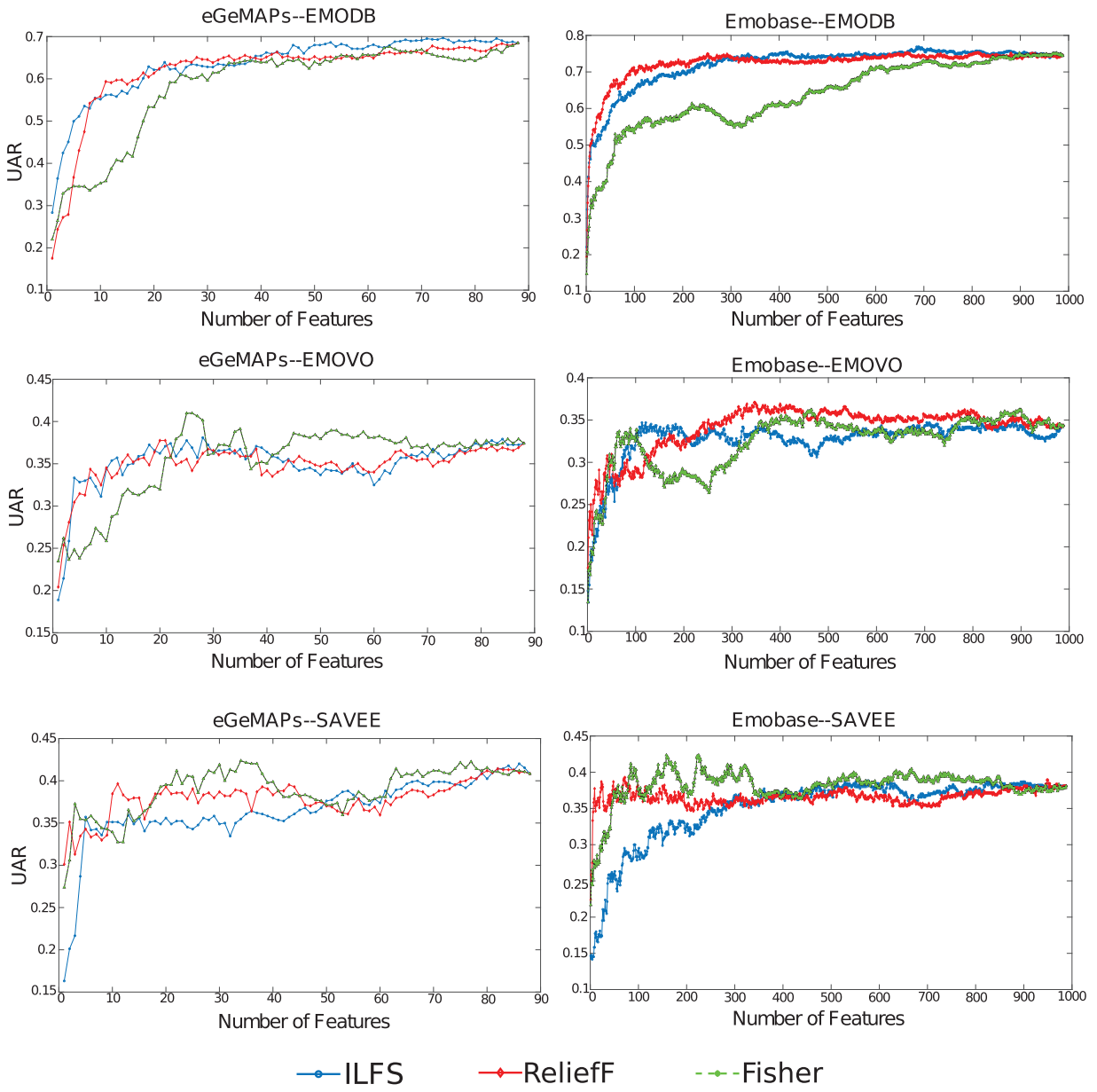


**Fig. 2.** Feature selection methods (ILFS, ReliefF and Fisher) results for all three data sets (EMODB, EMOVO and SAVEE) using two feature sets (eGeMAPs and emobase). Where x-axis represents the number of features and y-axis represents the UAR.

**Table 3**
Best Unweighted Average Recall (UAR (%)) of feature selection methods and number of selected features (numFeat) are reported. The best UAR (%) results for each feature set are given in bold. The unweighted arithmetic average for each feature selection method is also reported in 'Mean' column.

| Data Set | EmoDB | | | | EMOVO | | | | SAVEE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature Set | eGeMAPs | | emobase | | eGeMAPs | | emobase | | eGeMAPs | | emobase | | |
| | num Feat | UAR (%) | num Feat | UAR (%) | num Feat | UAR (%) | num Feat | UAR (%) | num Feat | UAR (%) | num Feat | UAR (%) | Mean |
| Baseline | 88 | 68.5 | 988 | 74.6 | 88 | 37.4 | 988 | 34.4 | 88 | 40.8 | 988 | 38.1 | 49.0 |
| ILFS | 74 | **69.7** | 685 | **76.9** | 28 | 38.1 | 113 | 34.7 | 86 | 42.0 | 574 | 38.8 | 46.9 |
| reliefF | 88 | 68.5 | 666 | 75.3 | 20 | 37.8 | 348 | **37.1** | 82 | 41.4 | 72 | 39.3 | 49.9 |
| Fisher | 88 | 68.5 | 975 | 75.2 | 25 | **41.0** | 464 | 36.2 | 34 | **42.4** | 158 | **42.4** | **51.0** |
| AFS | 81 | 68.5 | 696 | 75.8 | 2 | 39.0 | 56 | 36.4 | 68 | 40.5 | 21 | 37.5 | 49.6 |

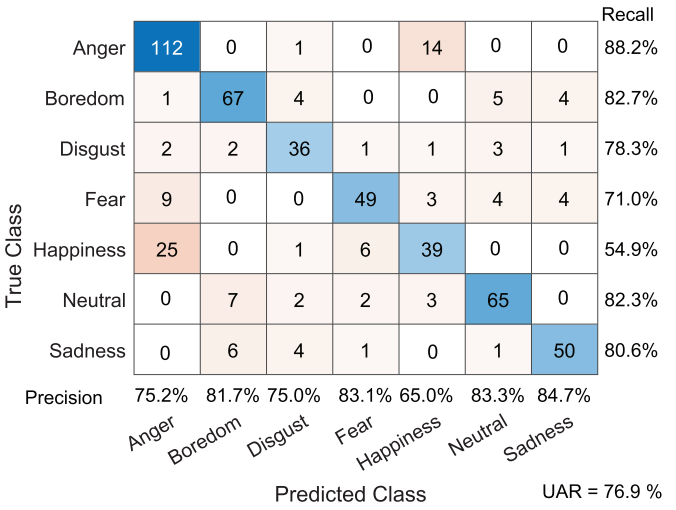


**Fig. 3.** Confusion matrix of ILFS Feature selection method for EmoDB data set using emobase feature set.

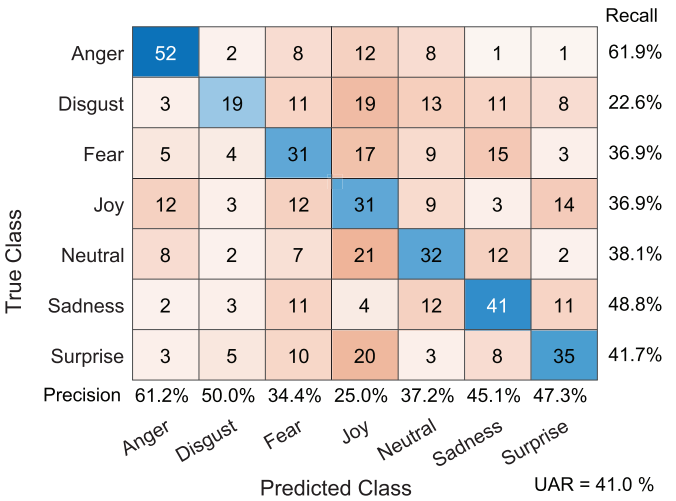


**Fig. 4.** Confusion matrix of Fisher feature selection method for EMOVO data set using eGeMAPs feature set.

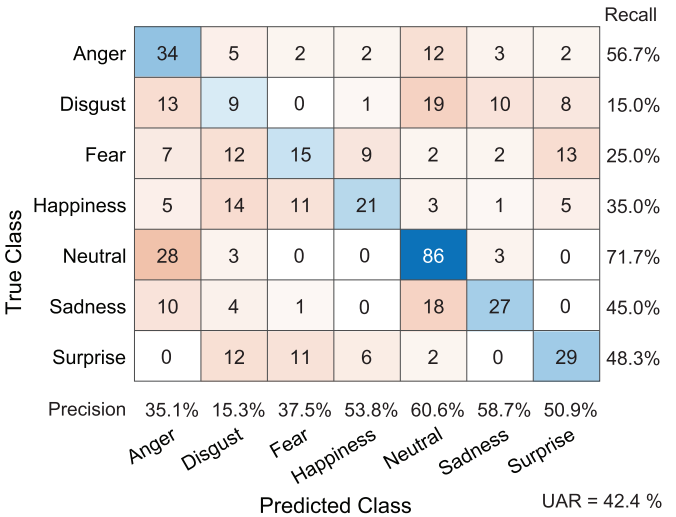


**Fig. 5.** Confusion matrix of Fisher feature selection method for SAVEE data set using eGeMAPs feature set.

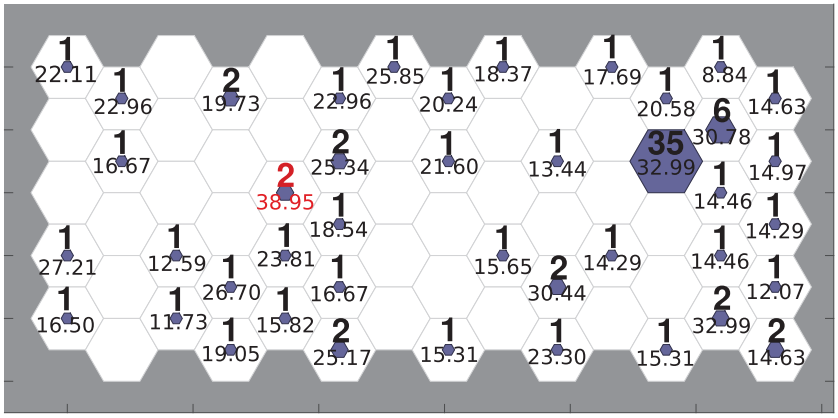


**Fig. 6.** A visualization of AFS method results: number of features present in each cluster (i.e. hexagon or neuron) along with the UAR (%) obtained using eGeMAPs feature set for EMOVO data set. Note that 2 out of 88 features provide better results than other feature subsets.

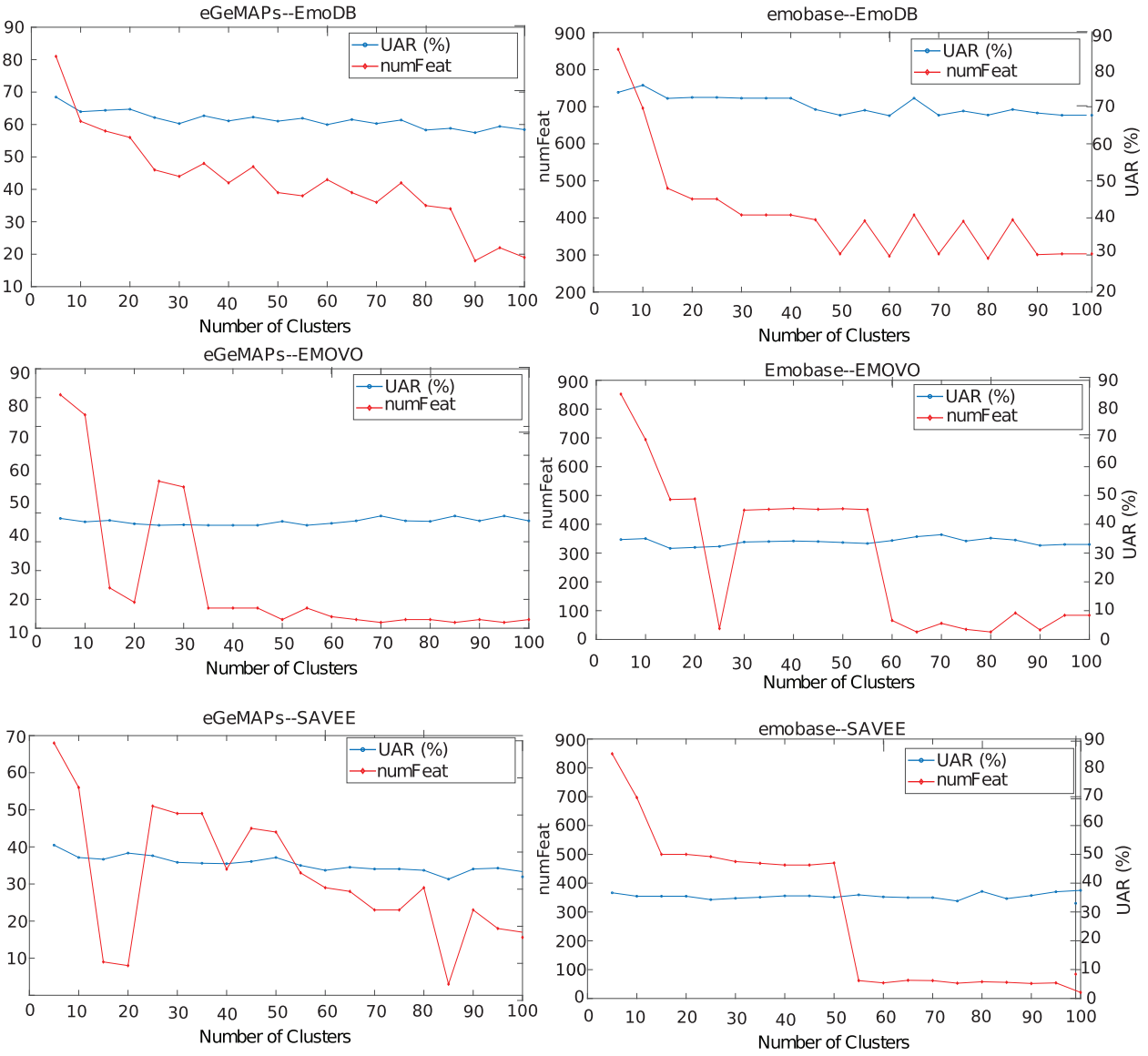


**Fig. 7.** AFS method results: The x-axis represents the number of cluster (N = 5,10,15, ... 100). The y-axis represents the number of features (numFeat) and Unweighted Average Recall (UAR) in % of the best cluster.

**Table 4**
Evaluation of feature selection methods for 7+1 emotion recognition task by combining all three data sets. Best Unweighted Average Recall (UAR (%)) and number of selected features (numFeat) are reported. The bold figures indicate the best UAR (%) for each feature set (i.e. eGeMAPs and emobase).

| Method | eGeMAPs | | emobase | |
|---|---|---|---|---|
| | numFeat | UAR (%) | numFeat | UAR (%) |
| Baseline | 88 | 44.4 | 988 | 47.4 |
| ILFS | 78 | 45.6 | 709 | 47.9 |
| relifF | 44 | **46.6** | 732 | **48.0** |
| Fisher | 53 | 45.3 | 822 | 47.9 |
| AFS | 79 | 43.8 | 835 | 47.2 |

To further evaluate the feature selection methods, we have combined all three data sets which results in a 8-class problem i.e. to recognise (7+1) emotions. The results of this experimentation in LOSO cross-validation setting is shown in Table 4. We have noted that the reliefF method provides the best results for eGeMAPs (46.6%) and emobase (48.0%) feature sets. All three data sets belong to different languages and have different qualities of annotation. Hence, the reliefF method could be a better choice than other methods where the quality and language of data sets are different.

In a previous study (Haider et al., 2018b), we demonstrated that the AFS method is able to select a feature subset which provides better results than the entire feature set and the PCA feature set for eating condition recognition. However the results have not been demonstrated in detail as in this study, and the AFS method has not been evaluated on multiple data sets and compared against other feature selection methods to the same extent as in this paper. The present study is therefore a step towards in demonstrating the generalisability of the AFS method. The contribution of this study is not only the evaluation of performance of different feature selection methods but also the assessment of the extent to which AFS, reliefF, Fisher and ILFS can reduce the feature set and therefore select small enough subsets which will impose lower computational demands on low resource systems, while preserving or improving emotion recognition performance, in comparison to full feature sets.

## 6. Conclusion

This study evaluated three state-of-the-art feature selection methods, namely, ILFS, reliefF and generalized Fisher score for emotion recognition, along with the recently proposed AFS method. It employed three different emotion recognition data sets from three different languages. The results show that higher UAR can be achieved using reduced feature sets. Generally, around 30 out of 88 eGeMAPs and 100 out of 988 emobase features are sufficient to obtain almost the same UAR as a full feature set. The Fisher feature selection method provided the best averaged UAR across all three data sets (51.0%) and two feature sets compared to the 49.0% averaged UAR for the full feature set baseline. However the reliefF method outperformed the other methods when all the data sets were combined. These findings are relevant to the development of machine learning models for machines with low computational resources. The AFS method provides competitive results in relation to the state of the art in feature select. AFS currently uses only features present in one cluster. For future studies, we will explore methods to rank the clusters of features and do fusion of different clusters for possible accuracy improvements. Other possible avenues for future work include testing the AFS on other modalities in addition to speech.

## Acknowledgement

This research is funded by the European Union's Horizon 2020 research program, under grant agreement No 769661, towards the SAAM project. PA is supported by the Medical Research Council (MRC). The work of S. Pollak was partially supported by the Slovenian Research Agency (ARRS) core research programme *Knowledge Technologies* (P2-0103).

## References

Aher, P.K., Daphal, S.D., Cheeran, A.N., 2016. Analysis of feature extraction techniques for improved emotion recognition in presence of additive noise. In: Proceedings of the International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). IEEE, pp. 350–354.

Akira, H., Haider, F., Cerrato, L., Campbell, N., Luz, S., 2015. Detection of cognitive states and their correlation to speech recognition performance in speech-to-speech machine translation systems. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association. International Speech Communications Association, pp. 2539–2543.

Anagnostopoulos, C.-N., Iliou, T., Giannoukos, I., 2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif. Intell. Rev. 43 (2), 155–177.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., 2005. A database of german emotional speech. In: Proceedings of the ninth European Conference on Speech Communication and Technology, pp. 1516–1520.

Consedine, N.S., Moskowitz, J.T., 2007. The role of discrete emotions in health outcomes: acritical review. Appl. Prevent. Psychol. 12 (2), 59–75.

Costantini, G., Iaderola, I., Paoloni, A., Todisco, M., 2014. Emovo corpus: an italian emotional speech database. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC). European Language Resources Association (ELRA), pp. 3501–3504.

Desmet, B., Hoste, V., 2013. Emotion detection in suicide notes. Expert Syst. Appl. 40 (16), 6351–6358.
Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T., 2017. From individual to group-level emotion recognition: Emotiw 5.0. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI). Association for Computing Machinery, pp. 524–528.
Dhall, A., Kaur, A., Goecke, R., Gedeon, T., 2018. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 2018 on International Conference on Multimodal Interaction. ACM, pp. 653–656.
Dimsdale, J.E., 2008. Psychological stress and cardiovascular disease. J. Am. Coll. Cardiol. 51 (13), 1237–1246.
El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. 44 (3), 572–587.
Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., et al., 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans. Affect. Comput. 7 (2), 190–202.
Eyben, F., Weninger, F., Groß, F., Schuller, B., 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia. ACM. Association for Computing Machinery, pp. 835–838.
Eyben, F., Wöllmer, M., Schuller, B., 2009. Openear–introducing the munich open-source emotion and affect recognition toolkit. In: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII). IEEE, pp. 1–6.
Goldshtein, E., Tarasiuk, A., Zigel, Y., 2011. Automatic detection of obstructive sleep apnea using speech signals. IEEE Trans. Biomed. Eng. 58 (5), 1373–1382.
Gu, Q., Li, Z., Han, J., 2012. Generalized fisher score for feature selection. arXiv preprint arXiv:1202.3725.
Haider, F., A. Salim, F., Conlan, O., Luz, S., 2018. An active feature transformation method for attitude recognition of video bloggers. In: Proc. Interspeech 2018, pp. 431–435. https://doi.org/10.21437/Interspeech.2018-1222.
Haider, F., Cerrato, L., Campbell, N., Luz, S., 2016. Presentation quality assessment using acoustic information and hand movements. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Institute of Electrical and Electronics Engineers (IEEE), pp. 2812–2816.
Haider, F., Cerrato, L.S., Luz, S., Campbell, N., 2016. Attitude recognition of video bloggers using audio-visual descriptors. In: Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction. Association for Computing Machinery, pp. 38–42.
Haider, F., de la Fuente, S., Luz, S., 2020. An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. IEEE J. Sel. Top. Signal Process. 14 (2), 272–281.
Haider, F., De La Fuente Garcia, S., Albert, P., Luz, S., 2020. Affective speech for alzheimer's dementia recognition. In: Kokkinakis, D., Lundholm Fors, K., Themistocleous, C., Antonsson, M., Eckerström, M. (Eds.), LREC: Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID). European Language Resources Association (ELRA), pp. 67–73.
Haider, F., Pollak, S., Zarogianni, E., Luz, S., 2018. SAAMEAT: active feature transformation and selection methods for the recognition of user eating conditions. In: Proceedings of the 2018 International Conference on Multimodal Interaction (ICMI). ACM. Association for Computing Machinery, pp. 564–568.
Hall, M.A., 1999. Correlation-based feature selection for machine learning. The University of Waikato Ph.D. thesis.
Haq, S., Jackson, P., 2009. Speaker-dependent audio-visual emotion recognition. In: Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP), pp. 53–58.
Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y., 2017. Learning supervised scoring ensemble for emotion recognition in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI). Association for Computing Machinery, pp. 553–560.
Jagini, N.P., Rao, R.R., 2017. Exploring emotion specific features for emotion recognition system using pca approach. In: Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, pp. 58–62.
Kira, K., Rendell, L.A., et al., 1992. The feature selection problem: Traditional methods and a new algorithm. Aaai, 2, pp. 129–134.
Knyazev, B., Shvetsov, R., Efremova, N., Kuharenko, A., 2017. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. arXiv preprint arXiv:1711.04598.
Kohonen, T., 1998. The self-organizing map. Neurocomputing 21 (1–3), 1–6.
Kononenko, I., Šimec, E., Robnik-Šikonja, M., 1997. Overcoming the myopia of inductive learning algorithms with ReliefF. Appl. Intell. 7 (1), 39–55.
Knig, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H., 2015. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. Alzheimer's Dementia 1 (1), 112–124.
Madzlan, N.A., Huang, Y., Campbell, N., 2015. Automatic classification and prediction of attitudes: audio-visual analysis of video blogs. In: Proceedings of the International Conference on Speech and Computer. Springer, pp. 96–104.
Mano, L.Y., Faial, B.S., Nakamura, L.H., Gomes, P.H., Libralon, G.L., Meneguete, R.I., Geraldo Filho, P.R., Giancristofaro, G.T., Pessin, G., Krishnamachari, B., 2016. Exploiting IoT technologies for enhancing health smart homes through patient identification and emotion recognition. Comput. Commun. 89, 178–190.
Ouyang, X., Kawaai, S., Goh, E.G.H., Shen, S., Ding, W., Ming, H., Huang, D.-Y., 2017. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI). ACM. Association for Computing Machinery, pp. 577–582.
Robnik-Šikonja, M., Kononenko, I., 1997. An adaptation of Relief for attribute estimation in regression. In: Proceedings of the Fourteenth International Conference on Machine Learning (ICML), ICML 1997, 5, pp. 296–304.
Roffo, G., Melzi, S., Castellani, U., Vinciarelli, A., 2017. Infinite latent feature selection: A probabilistic latent graph-based ranking approach. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 1407–1415.
Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. 53 (9–10), 1062–1087.
Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y., 2014. The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load. In: Proceedings of the 15th Annual conference of the International Speech Communication Association. International Speech Communications Association, pp. 427–431.
Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hnig, F., Orozco-Arroyave, J.R., Nth, E., Zhang, Y., Weninger, F., 2015. The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association, pp. 478–482.
Su, J., Luz, S., 2016. Predicting cognitive load levels from speech data. Recent Advances in Nonlinear Speech Processing. Springer, pp. 255–263.
Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M., 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (AVEC). Association for Computing Machinery, pp. 3–10.
Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features, and methods. Speech Commun. 48 (9), 1162–1181.
Vielzeuf, V., Pateux, S., Jurie, F., 2017. Temporal multimodal fusion for video emotion classification in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI). Association for Computing Machinery, pp. 569–576.
Wang, J., Chang, C.-I., 2006. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. IEEE Trans. Geosci. Remote Sens. 44 (6), 1586–1600.
Wang, S., Ling, X., Zhang, F., Tong, J., 2010. Speech emotion recognition based on principal component analysis and back propagation neural network. In: Proceedings of the International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 3. IEEE, pp. 437–440.
Wang, S., Wang, W., Zhao, J., Chen, S., Jin, Q., Zhang, S., Qin, Y., 2017. Emotion recognition with multimodal features and temporal models. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI). Association for Computing Machinery, pp. 598–602.
Weninger, F., Eyben, F., Schuller, B.W., Mortillaro, M., Scherer, K.R., 2013. On the acoustics of emotion in audio: what speech, music, and sound have in common. Front. Psychol. 4.