SAAMEAT: Active Feature Transformation and Selection Methods for the Recognition of User Eating Conditions

Fasih Haider
Usher Institute of Population Health Sciences & Informatics
Edinburgh Medical School
The University of Edinburgh
Edinburgh, UK
Fasih.Haider@ed.ac.uk

Senja Pollak
Usher Institute of Population Health Sciences & Informatics
Edinburgh Medical School
The University of Edinburgh
Edinburgh, UK
senja.pollak@ed.ac.uk

Eleni Zarogianni
Usher Institute of Population Health Sciences & Informatics
Edinburgh Medical School
The University of Edinburgh
Edinburgh, UK
e.zarogianni@ed.ac.uk

Saturnino Luz
Usher Institute of Population Health Sciences & Informatics
Edinburgh Medical School
The University of Edinburgh
Edinburgh, UK
s.luz@ed.ac.uk

ABSTRACT

Automatic recognition of eating conditions of humans could be a useful technology in health monitoring. The audio-visual information can be used in automating this process, and feature engineering approaches can reduce the dimensionality of audio-visual information. The reduced dimensionality of data (particularly feature subset selection) can assist in designing a system for eating conditions recognition with lower power, cost, memory and computation resources than a system which is designed using full dimensions of data. This paper presents Active Feature Transformation (AFT) and Active Feature Selection (AFS) methods, and applies them to all three tasks of the ICMI 2018 EAT Challenge for recognition of user eating conditions using audio and visual features. The AFT method is used for the transformation of the Mel-frequency Cepstral Coefficient and *ComParE* features for the classification task, while the AFS method helps in selecting a feature subset. Transformation by Principal Component Analysis (PCA) is also used for comparison. We find feature subsets of audio features using the AFS method (422 for Food Type, 104 for Likability and 68 for Difficulty out of 988 features) which provide better results than the full feature set. Our results show that AFS outperforms PCA and AFT in terms of accuracy for the recognition of user eating conditions using audio features. The AFT of visual features (facial landmarks) provides less accurate results than the AFS and AFT sets of audio features. However, the weighted score fusion of all the feature set improves the results.

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10.
https://doi.org/10.1145/3242969.3243685

CCS CONCEPTS

• **Information systems → Multimedia databases**; • **Computing methodologies → Speech recognition**; **Feature selection**;

KEYWORDS

feature transformation, feature selection, audio-visual processing, feature extraction, eating condition, dimensionality reduction

ACM Reference Format:
Fasih Haider, Senja Pollak, Eleni Zarogianni, and Saturnino Luz. 2018. SAAMEAT: Active Feature Transformation and Selection Methods for the Recognition of User Eating Conditions. In *ICMI '18: 2018 Int'l Conference on Multimodal Interaction, Oct. 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3242969.3243685

1 INTRODUCTION

Developing multimodal systems for recognizing food and eating-related conditions in a speaking context can contribute to several research areas, including automatic speech recognition and speaker identification [5]. Another important application area may be health monitoring. As our diet has an impact on our health and well-being [1], automatically recognizing eating-related events could contribute to the development of automated and non-obtrusive health monitoring technologies. Similarly, ambient and assisted living technologies which focus on monitoring of everyday activities could benefit from recognition of eating activity.

The ICMI 2018 Eating Analysis and Tracking (EAT) Challenge addresses the audio-visual classifications of user eating conditions, with three sub-challenges. In the *Food-type Sub-Challenge* the task is to classify an utterance into one of the seven food types that the subject is eating while speaking (one of the being no eating, and the other six classes are Apple, Nectarine, Banana, Crisp, Biscuit and Gummi-bear). In the *Likability Sub-Challenge* the aim is to recognize the subjects' food likability rating, and the *Chew and Speak*

Difficulty Sub-Challenge concerns the level of difficulty to speak while eating. The dataset used is iHEARu-EAT database [5], which partly featured also as the data for the Interspeech 2015 ComParE Challenge [12]. The winning approach of the Interspeech challenge conducted on iHEARu-EAT used the Fisher vector encoding of MFCCs and PLP features for the recognition of food type using extreme learning machines and partial least square regression based classifiers, which reported an Unweighted Average Recall (UAR) of 83.1% on the test dataset [8]. Although the results are promising but no effort is spent on dimensionality reduction (removing noisy/redundant features) to reduce the 'curse of dimensionality' and computational resources (i.e. extraction of a subset of feature set instead of whole feature set results in reduction of usage of machine memory and computational power). The current state of the art methods for user eating conditions recognition may get benefits from the dimensionality reduction methods for further improvement in accuracy and for the situations where the computational and memory resources are limited.

In this paper, we use a recently developed Active Feature Transformation (AFT) method [3] and propose a novel method for active feature selection (AFS). The main contributions of this paper are:

(1) Introduction of a novel feature selection method, namely AFS, and demonstration of the discrimination power of acoustic (*emobase* feature set) features and their Principle Component Analysis (PCA) representation.
(2) Evaluation of active feature selection and transformation methods against feature without transformation/selection and EAT-Baselines in terms of accuracy and dimensionality.
(3) Demonstration of discrimination power of MFCC features for the EAT challenge.
(4) Demonstration of discrimination power of AFT applied to the *EAT challenge's* audio-visual features (*ComParE* feature set and normalized-facial landmarks).

The paper is structured as follows. In Section 2 we introduce the active feature transformation method, followed by active feature selection in Section 3. After introducing the ICMI 2018 EAT dataset in Section 4, we describe the experiments (in Section 5) and the results (in Section 6) of all three challenges using audio-visual features. In Section 7 we conclude the paper, highlight the main contributions and present the plans for future work.

2 ACTIVE FEATURE TRANSFORMATION

The AFT method for MFCC features has been recently proposed [3] to detect the attitudes of video bloggers, and showed significant improvement in terms of dimensionality and accuracy over PCA and MFCC features. The AFT method consists of the following steps for feature transformation:

(1) First a speech segment (Si) is divided into n frames ($F_{k,Si}$) of fixed duration (100 ms) with no overlap with the neighboring frame, where $i = 1 : N$ and N represents the total number of speech segments, and $k = 1 : n$, that is k varies from 1 to n, the total number of frames in a speech segment (Si). Hence $F_{k,Si}$ is the k^{th} frame of i^{th} speech segment, and 228 MFCC features are extracted over a frame ($F_{k,Si}$), rather than over speech segments of variable duration. The system architecture is depicted in Figure 1.
(2) Clustering of frames: We used Self-organizing Maps (SOM) [9] for the clustering of frames into n clusters ($C1, C2, ..., Cn$), as depicted in Figure 2 (n represents the cluster size for SOM).
(3) Generation of an active feature transformation (AFT_{Si}) vector by calculating the number of frames in each cluster for each speech segment (Si) as depicted in Figure 2.
(4) As the number of frames are different for each speech segment (i.e. the duration of each speech segment is not constant), we normalize the feature vector by dividing it by the total number of frames present in each speech segment ($\sum AFT_{Si}$) as set out in Equation 1.

$$AFT_{Si_{norm}} = \frac{AFT_{Si}}{\sum AFT_{Si}} \quad (1)$$

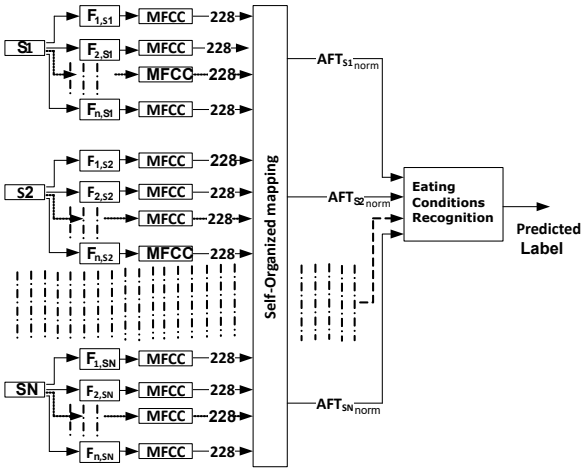

Figure 1: Active feature transformation method applied to MFCC features.

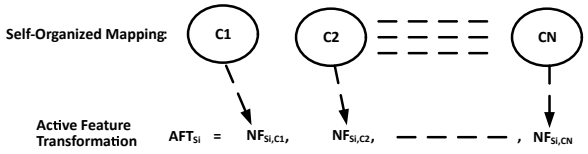

Figure 2: Active Feature Transformation: AFT_{Si} represents active feature transformation of i^{th} speech segment and $NF_{Si,Ck}$ represent total number of frames of i^{th} speech segment within k^{th} cluster. Where $k = 1 : N$ and N is the total number of clusters.

3 ACTIVE FEATURE SELECTION

In this section, we describe our active feature selection method which divides a feature set into subsets. It involves clustering the dataset into N (where $N = 5, 10, 15, ..., 100$) clusters using self-organizing maps, and then evaluating features present in each cluster through a Leave One Subject Out (LOSO) cross-validation setting, as depicted in Figure 3, and selecting the cluster with the highest result. Here, we are not clustering the number of instances but the dimensions. Our hypothesis is that the noisy features have

different characteristics than informative features, and that clustering the features will divide the features into many subsets according to their common characteristics. An example of self-organizing clustering is depicted in Figure 4, where 988 features (*emobase.conf*) are clustered into 30 clusters (feature subsets). The distance between these clusters is depicted in Figure 5.

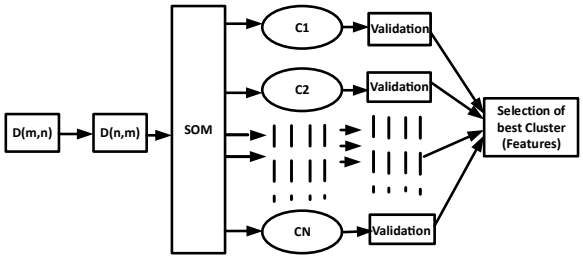

Figure 3: Active feature selection method: D(m,n) represents the data where m is the total number of instances (in EAT challenge case it is 945 instances of train data set) and n is the total number of dimensions (988 openSmile features extracted using the *emobase.conf* file). The validation is performed in Leave One Subject Out (LOSO) cross-validation setting.

4 DATASET

We used both audio and video recordings of the challenge, taken from the IHEARuEAT database [5]. This dataset consists of recordings taken from 30 German-speaking subjects (15 males, 15 females, mean age 26.1). Prior to the recording of each utterance, a participant was provided with a serving (the serving size was personalized so that it had a significant effect on the subject's speech, while ensuring that the production of speech was audible). Food classes were selected for partly similar consistency (e.g., crisps and biscuits) and partly dissimilar consistency (e.g., nectarine and crisps) and for being also frequently encountered in practice. The recorded data comprised of both read and spontaneous speech.

Participants assessed how much they liked the food they were eating during the experiment (continuous slider between 0-dislike extremely to 1-like extremely). A 5-point Likert scale was also used after the recording, to specify any difficulties the participants have encountered in eating each type of food while speaking. The data was split into a training set of 20 and a test set of 10 speakers, stratified by age and gender.

5 EXPERIMENTATION

5.1 Feature Extraction

The features vectors for the different experiments are as follows:

(1) **Experiment One (*emobase*):** 988 acoustic features extracted over each speech segment (video clip) using *emobase.config* configuration file of openSMILE [2] which has been widely used for emotion recognition [10]. The feature set includes spectral parameters (MFCC, LSP, etc.), intensity and pitch related parameter.
(2) **Experiment Two ($PCA_{emobase}$):** Principle component analysis of experiment one features.
(3) **Experiment Three ($mfcc$):** 228 MFCC features (a subset of the feature set used in Experiment one) extracted for each speech segment. OpenSMILE calculates an overall MFCC feature response for speech segments with variable duration using statistical functionals such as mean, standard deviation, minimum, maximum, range values etc. The objective of calculating an overall response is to project the features onto a fixed number of dimensions (in this case 228) for machine learning methods (e.g. dimensionality reduction and classification).
(4) **Experiment Four (AFT_{mfcc}):** Transformed version of 228 MFCC features, which are extracted for a frame of fixed duration (100 ms) instead of full speech segment using AFT as described in Section 2.
(5) **Experiment Five ($AFT_{Compare}$):** Transformed version of *ComParE* features (provided by Challenge Organizers) using AFT.
(6) **Experiment Six (*AFS*):** Evaluation of Active Feature Selection method using the experiment one features.
(7) **Experiment Seven (AFT_{Visual}):** Transformed version of normalized visual landmarks features (provided with the EAT Challenge data) using AFT.
(8) **Experiment Eight (*Fusion*):** Weighted score fusion of all the above experiments using the weights depicted in Table 1. The weights are calculated using brute force approach for two classifiers and then the best resulted score is fused with third classifier's score, the weights are searched again using brute force approach for the best fused score of two classifier and third classifier and so on.

5.2 Classification Methods and Evaluation Measures

The classification is performed using Linear Discriminant Analysis (LDA). This classifier is employed in MATLAB[1] using the statistics and machine learning toolbox in the Leave one subject out (LOSO) cross-validation setting. LDA works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix [11]).

For evaluation, we use Unweighted Average Recall (UAR) for the 'Food Type' and 'Likability' Challenges, which is the average of the recall of all the classes, and we use the Concordance Correlation Coefficients (CCC) for the difficulty Challenge, following [4].

6 RESULTS AND DISCUSSION

The classification is performed in a LOSO (Leave one subject out) cross-validation setting using the feature vectors described in Section 5.1 and the best results of all the experiments and number of dimensions used for classification are summarized in Table 2. It is observed that the *AFS* method provides the best results for all three challenges, and also outperforms the *End2You* baseline method [4]. However, it is unable to outperform the *openXBOW* baseline method [4]. The subsets of features selected using *AFS* provide better results than full feature set. These results validated our hypothesis that by clustering the features we can remove noisy/redundant

[1] https://uk.mathworks.com/products/matlab.html (August 2018)

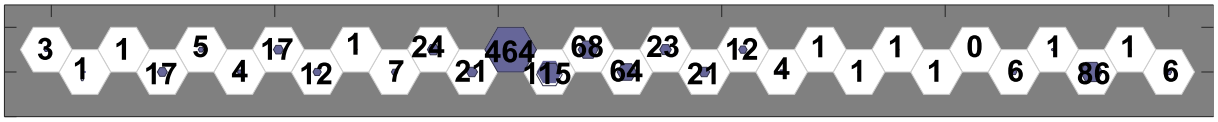

Figure 4: Figure indicates the number of features present in each cluster. Where N = 30 and the cluster with highest accuracy contain 68 out of 988 features for the 'Difficulty Challenge'.

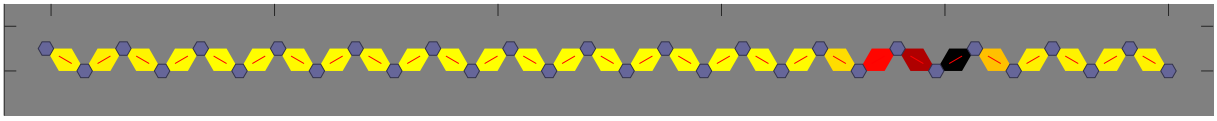

Figure 5: Figure indicates the distance between clusters (blue dots) and darker color indicates greater distance between clusters. Where features in Cluster number 16 (68 out of 988) provides the best validation results for the 'Difficulty Challenge'.

Table 1: weighted Score fusion of experiments.

$$\begin{cases} Fusion_{FoodType} = 0.00.emobase + 0.14.PCA_{emobase} + 0.137.mfcc + 0.17.AFT_{mfcc} + 0.24.AFT_{Compare} + 0.27.AFS + 0.04.AFT_{Visual} \\ Fusion_{Likability} = 0.05.emobase + 0.00.PCA_{emobase} + 0.00.mfcc + 0.25.AFT_{mfcc} + 0.18.AFT_{Compare} + .0.33.AFS + 0.20.AFT_{Visual} \\ Fusion_{Difficulty} = 0.01.emobase + 0.087.PCA_{emobase} + 0.07.mfcc + 0.12.AFT_{mfcc} + 0.16.AFT_{Compare} + 0.21.AFS + 0.35.AFT_{Visual} \end{cases}$$

Table 2: Best results for each experiment in LOSO cross-validation setting on train dataset along with dimensionality of features used for classification: Food Type and Likability results are in UAR and dimensionality (UAR,dimensionality), and Difficulty in CCC and dimensionality (CCC, dimensionality). Where only AFT_{Visual} uses the visual features and the rest of experiments use the acoustic features. The OPENXBOX and EndtoYOU are baseline methods.

	$emobase$	$PCA_{emobase}$	$mfcc$	AFT_{mfcc}	$AFT_{Compare}$	AFS	AFT_{Visual}	Fusion	*openXBOW*	*EndtoYOU*
Food Type	29.17%, 988	51.72%, 481	43.11%, 288	39.30%, 40	33.59%, 65	*54.81%*, 422	23.84%, 80	**59.35%**	***64.30%***	35.20%
Likability	55.73%, 988	60.46%, 296	61.94%, 288	64.12%, 40	62.60%, 95	*65.50%*, 401	61.12%, 70	***68.99%***	**66.50%**	55.10%
Difficulty	0.059, 988	0.372, 171	0.295, 288	0.322, 55	0.242, 65	*0.418*, 68	0.281, 85	**0.470**	***0.481***	0.345

features, as by clustering (AFS method) the total number of *emobase* features, we find a subset of features (422 for Food Type, 104 for Likability and 68 (as depicted in Figure 4) for Difficulty, out of 988 features) which provide better results than the full feature set (*emobase*). However, it is not explicitly clear what common characteristic a cluster of features has. One of the possible explanation is that as we are validating in LOSO setting (evaluating each cluster) for the tasks, the cluster with higher validation UAR may contain features which have a common characteristic of speaker in-dependency than other clusters for the tasks. The score fusion of all the feature sets improves the results for all sub-challenges.

The $AFT_{Compare}$ provides less accurate results than AFT_{mfcc}, suggesting that the MFCC features may improve the performance of *openXBOW* method. The *AFS* provides promising results and also suggests that the selected feature should be extracted over frame level and then feed into *openXBOW* method. The *ComParE* features set used for *openXBOW* and $AFT_{Compare}$ method is calculated for each 10 ms and MFCC features for AFT_{mfcc} are extracted over 100 ms. The different frame size should also be explored for both *openXBOW* and AFT method as the speech signal properties varies over time. Besides that the features reported in this study may improve the performance of *openXBOW* method. However, the *AFS* is not able to outperform the previous challenge methods but clearly demonstrate that the higher accuracy can be achieved with a lower dimension of feature than full feature set. The transformed and selected subsets of features should be tested with other advanced classifiers such as extreme learning machines [6, 7] and partial least squares [13].

7 CONCLUSIONS

A novel Active Feature Selection (AFS) method has been proposed and used for the recognition of eating conditions. The subset of features selected using the AFS outperformed the full feature set and the PCA transformation of the full feature set. However, the results reported in this paper do not outperform the baseline of *openXBOW* method (though they do outperform the *End2You* method). Possible extensions of the work presented here include testing these methods in conjunction with advanced classifiers to assess whether any possible improvements in accuracy, and generating Fisher vectors using the subset of features selected by AFS. In future we intend to evaluate the performance of the AFT and AFS methods for multiple feature sets, including prosodic, voice quality, and image features on multiple prediction problems such as sound events detection, emotion recognition, human action recognition and autism recognition.

ACKNOWLEDGMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661, SAAM project.

REFERENCES

[1] Nathalie T. Burkert, Johanna Muckenhuber, Franziska Großschädl, Éva Rásky, and Wolfgang Freidl. 2014. Nutrition and Health – The Association between Eating Behavior and Various Health Parameters: A Matched Sample Study. *PLoS ONE* 9, 2 (feb 2014), e88278. https://doi.org/10.1371/journal.pone.0088278

[2] Florian Eyben, Felix Weninger, Florian Groß, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[3] Fasih Haider, Fahim Salim, Owen Conlan, and Saturnino Luz. 2017. An Active Feature Transformation Method For Attitude Recognition of Video Bloggers. In *Proc. Interspeech 2017*.

[4] Simone Hantke, Maximilian Schmitt, Panagiotis Tzirakis, and Björn Schuller. 2018. EAT - The ICMI 2018 Eating Analysis and Tracking Challenge. In *Proceedings of the 2018 ACM on International Conference on Multimodal Interaction*. ACM.

[5] Simone Hantke, Felix Weninger, Richard Kurle, Fabien Ringeval, Anton Batliner, Amr El-Desoky Mousa, and Björn Schuller. 2016. I Hear You Eat and Speak: Automatic Recognition of Eating Condition and Food Type, Use-Cases, and Impact on ASR Performance. *PLOS ONE* 11, 5 (may 2016), e0154486. https://doi.org/10.1371/journal.pone.0154486

[6] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. 2012. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42, 2 (2012), 513–529.

[7] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. 2006. Extreme learning machine: theory and applications. *Neurocomputing* 70, 1-3 (2006), 489–501.

[8] Heysem Kaya, Alexey A Karpov, and Albert Ali Salah. 2015. Fisher vectors with cascaded normalization for paralinguistic analysis. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[9] Teuvo Kohonen. 1998. The self-organizing map. *Neurocomputing* 21, 1-3 (1998), 1–6.

[10] Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen. 2014. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 494–501.

[11] Sarunas Raudys and Robert P. W. Duin. 1998. Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters* 19, 5-6 (April 1998), 385–392.

[12] Björn W. Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. 2015. The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*. 478–482. http://www.isca-speech.org/archive/interspeech_2015/i15_0478.html

[13] Herman Wold. 1985. Partial least squares. *Encyclopedia of statistical sciences* (1985).