Building Parsimonious SVM Models for Chewing Detection and Adapting Them to the User

Iason Karakostas, Vasileios Papapanagiotou(✉), and Anastasios Delopoulos

Multimedia Understanding Group, Department of Electrical
and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece
iasonekv@auth.gr, vassilis@mug.ee.auth.gr, adelo@eng.auth.gr
https://mug.ee.auth.gr

Abstract. Monitoring of eating activity is a well-established yet challenging problem. Various sensors have been proposed in the literature, including in-ear microphones, strain sensors, and photoplethysmography. Most of these approaches use detection algorithms that include machine learning; however, a universal, non user-specific model is usually trained from an available dataset for the final system. In this paper, we present a chewing detection system that can adapt to each user independently using active learning (AL) with minimal intrusiveness. The system captures audio from a commercial bone-conduction microphone connected to an Android smart-phone. We employ a state-of-the-art feature extraction algorithm and extend the Support Vector Machine (SVM) classification stage using AL. The effectiveness of the adaptable classification model can quickly converge to that achieved when using the entire available training set. We further use AL to create SVM models with a small number of support vectors, thus reducing the computational requirements, without significantly sacrificing effectiveness. To support our arguments, we have recorded a dataset from eight participants, each performing once or twice a standard protocol that includes consuming various types of food, as well as non-eating activities such as silent and noisy environments and conversation. Results show accuracy of 0.85 and F1 score of 0.83 in the best case for the user-specific models.

Keywords: Active learning · Dietary monitoring
Chewing detection · Wearable sensors

1 Introduction

Automatically monitoring eating activity has received significant attention in the research community; a variety of novel sensors and detection algorithms have been proposed that monitor eating activity based on detecting chewing or swallowing. Microphones are often used, placed either near the throat in order to detect swallowing sounds [8], or in-ear to detect chewing sounds [1,5]. More recent approaches rely on strain sensors to capture muscle activity [8,9] and

S. Battiato et al. (Eds.): ICIAP 2017 International Workshops, LNCS 10590, pp. 403–410, 2017.
https://doi.org/10.1007/978-3-319-70742-6_38

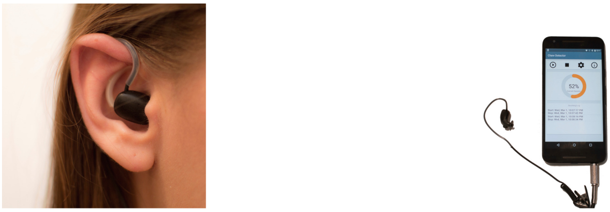
detect chewing. Other types of proposed sensors include custom build or modified proximity sensors placed on the wrists and head of the subject that detect hand movement transferring food from plate to mouth [3], or use commercial smartwatches to detect food-intake cycles [4].

Some systems take advantage of multiple sensor signals. In [7], a custom built sensor which consists of an open-air in-ear microphone, a photoplethysmography sensor, and an acceleromenter is proposed; the detection algorithm calculates, among other features, the fractal dimension of the microphone signal as proposed in [6]. Authors report accuracy of 0.938 and F1 score of 0.761. In [10], a system with two off-the-shelf in-ear bone-conduction microphones is proposed. Through spectrum analysis and k-NN classification the system can differentiate between eating, speaking and drinking activities, with average intra-subject accuracy of 0.8 and average inter-subject accuracy of 0.7.

Most of the proposed algorithms incorporate machine learning, usually the Support Vector Machine (SVM), and more recently convolutional neural networks [5]. Thus, they require sufficient data to train an effective classification model; this model can then be deployed in the final system, and is used unaltered during its life-span. However, not every person eats and chews in exactly the same manner; thus, it is reasonable for a system to be able to adapt to each different user to increase its effectiveness overall. In this paper, we propose such a user adaptable chewing detection system based on an off-the-self bone conduction in-ear microphone with an integrated speaker and an Android smart-phone. We employ the same audio features of [7] to detect eating events, and extend it with active learning (AL) for the SVM classification step in two-fold. First, we propose a non-interactive strategy to create a parsimonious SVM model that includes few support vectors (SVs) and is thus easy to compute and retrain on smart phones, and show that the effectiveness of the parsimonious model quickly converges to that of an equivalent SVM model. Second, we propose a feedback-request (interactive) method with low user intrusiveness for the deployed version of our system that enables per user-adaptation of the deployed model in order to increase effectiveness. The rest of this paper is organised as follows. Section 2 presents the bone-conduction microphone, the captured signals, and the extracted features. Section 3 presents the SVM-based chewing classification and AL. In Sect. 4 the experimental dataset is presented along with the evaluation results of the system. Finally, Sect. 5 concludes the paper.

2 Chewing Sensor and Signals

The chewing detection hardware consists of an off-the-shelf bone-conduction microphone connected via wire to the headphone jack of an Android smartphone (Fig. 1b). The microphone (Invisio M3h) exhibits a sensitivity of −32 dB around the 1 kHz frequency band. We sample audio at 4 kHz which is the lowest sampling frequency allowed by the Android operating system, in order to reduce the computational burden without sacrificing effectiveness [6]. The microphone is housed in an ear-bud which is placed inside the outer ear canal (Fig. 1).



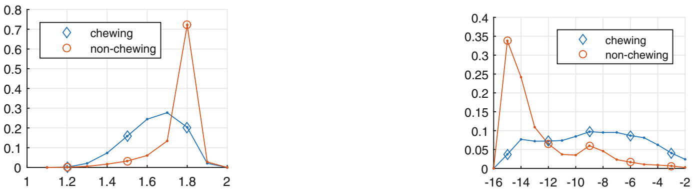
(a) The bone-conduction microphone

(b) The complete system

Fig. 1. The microphone placement, and the complete detection system.

We have opted for a bone-conduction microphone since it can naturally eliminate external (non-user generated) sounds very well, since sound is captured by measuring the vibration transmitted through user's bones, not air pressure.

The microphone captures body-generated sounds such as voice and chewing sounds. A pre-processing step applies a high-pass FIR filter with a cut-off frequency of 20 Hz in order to remove low frequency content, which does not include chewing-related information. Subsequently, we extract overlapping windows every 160 samples (sampling interval of $T_f = 25$ Hz) and compute a feature vector $\mathbf{f}[n]$ (where n is the time-index corresponding to nT_f sec). The set of extracted features are described in detail in [7] and in this work are augmented with the signal's variance, resulting in 16 features in total. The time-domain features (fractal dimension, moments) are computed on windows of 0.1 s (400 samples), while the spectral features on windows of 0.2 s (800 samples). Before computing the features, each window is normalised by subtracting its mean and dividing it by its standard deviation (with the exception of variance). Figure 2 shows the histograms of fractal dimension (2a) and log of variance (2b) for the chewing and non-chewing classes. Both features are quite discriminative individually; combining all the extracted features can better distinguish chewing from non-chewing windows. More details about each feature can be found in [7].



(a) Fractal dimension histogram

(b) Log of variance histogram

Fig. 2. Histograms of the fractal dimension (a) and log of variance (b) for the chewing and non-chewing classes.

3 Classification and Active Learning

In order to classify each feature vector $\mathbf{f}[n]$ into chewing (positive class) or non-chewing (negative class) we employ the SVM classifier with RBF kernel. Initially, each feature is smoothed in the time domain using a Hamming filter of 3.72 s. The SVM scores are computed as $s[n] = \mathbf{w} \cdot \mathbf{f}[n] + b$ where $\mathbf{w}$ is the separating hyper-plane normal vector, and b is the offset. Parameters C of SVM and γ of RBF are chosen based on preliminary experiments described in Sect. 4.1.

AL is a method of improving a classifier's effectiveness [11] by enhancing the training set in "rounds". In each round, the current classification model is applied on a pool of available feature vectors; some few feature vectors are selected from the pool and the user is requested to provide feedback (the correct label) for these feature vectors. As a result, it is not necessary to annotate the entire pool. In this work, we propose to use AL in two distinct tasks: (a) a parsimonious AL training (PALT) approach without any user feedback to build a reduced complexity classifier (small number of SVs), and (b) an inter-active learning adaptation (IALA) strategy to adapt a pre-trained model to the user's eating style and improve effectiveness.

3.1 PALT

Given a training set $\mathcal{T} = \{(\mathbf{f}[i], y[i]) : i = 0, 1, \ldots, N-1\}$, one can directly train an SVM model $\mathcal{M}$. PALT uses AL to create a model with much fewer SVs without sacrificing the model's discriminative power. First, a few items of $\mathcal{T}$ are selected and form the initial training set $\mathcal{T}_0$, and an SVM model $\mathcal{M}_0$ is trained on it. The model is applied on the remaining data $\mathcal{P}_0 = \mathcal{T} - \mathcal{T}_0$ and $s[i]$ is computed for each item of $\mathcal{P}_0$. We then select the l positive misclassifications ($s[i] > 0$ and $y[i] = -1$) that are closest to the separating hyperplane, and similarly the l negative misclassifications ($s[i] < 0$ and $y[i] = +1$) closest to the separating hyperplane, since such vectors are more likely to become SVs. Let $\mathcal{U}_0$ be the set of the selected $2l$ misclassifications. For the next round, we create the new training set as $\mathcal{T}_1 = \mathcal{T}_0 \cup \mathcal{U}_0$; a new SVM model $\mathcal{M}_1$ is trained on $\mathcal{T}_1$, and then applied to $\mathcal{P}_1 = \mathcal{P}_0 - \mathcal{U}_0$. A new AL round can take place, and this process can continue until there are no more misclassifications or the model is "large" or "effective enough", depending on the requirements for computational complexity and effectiveness. Thus, PALT collects the vectors that are more likely to become SVs and affect the separating hyperplane orientation the most.

In our case, we start with a $\mathcal{T}_0$ that contains 20 positive and 20 negative feature vectors, selected randomly from $\mathcal{T}$. An initial training set of 40 vectors is very small and allows us to observe the effect of augmenting the training set. At each AL round, we add only one positive and one negative misclassification ($l = 1$). We repeat this process for 800 rounds, so our final model is $\mathcal{M}_{800}$. In Sect. 4.1 we compare $\mathcal{M}_{800}$ to the model $\mathcal{M}$ obtained by training directly on the entire $\mathcal{T}$.

3.2 IALA

This method aims at adapting a pre-trained SVM model to a single user based on inter-active feedback requests for ambiguous time intervals. The pre-trained model can be that of a straight-forward SVM training, or the result of the PALT approach proposed in Sect. 3.1. Note that PALT cannot be directly applied in this scenario for two reasons. First, it requires a large dataset to work on, and thus precious storage space of smart phones. Second, for every feature vector that feedback is required, the user would have to recall past eating activity, however evidence shows that people tend to under-report eating [2]. This would compromise the feedback quality, rendering PALT useless.

IALA is based on two thresholds, a time threshold t_{thr} and an SVM score threshold s_{thr}, and detecting in real-time ambiguous intervals; the user is then immediately asked if she/he is eating. Let $\mathbf{f}[n]$ by the stream of feature vectors and $s[n]$ their SVM scores based on the current model $\mathcal{M}_c$, trained on $\mathcal{T}_c$. The most recent $\mathbf{f}[n]$ along with their $s[n]$ are buffered in memory, so that they cover the last t_{thr} seconds. If the SVM score is closer to the separating hyperplane than s_{thr} for all of the buffered vectors, the user is immediately asked if she/he is eating. The requirement for t_{thr} eliminates false alarms, and prevents the system from overwhelming the user with feedback requests and $\mathcal{T}_c$ with new vectors. From the buffered feature vectors, the q closest to the hyperplane are added in $\mathcal{T}_c$ with the label the user provides, and a new model is trained. Through the experiments we have set $t_{\mathrm{thr}} = 1\,\mathrm{s}$, $s_{\mathrm{thr}} = 0.2$, and $q = 6$, as we have observed that windows of 1 s where $s[n]$ remains $s_{\mathrm{thr}} = 0.25$ do not occur too often, so that the user is not constantly asked for feedback.

4 Experimental Evaluation

To evaluate our proposed system, we have collected a dataset of audio recordings using the bone-conduction microphone and a dedicated Android application that allows easy time-stamping. The experimentation protocol followed by each participant includes chewing activities with 7 different food types, as well as common non-chewing activities (e.g. walking, talking, and listening to music) both in silent and noisy setups. The dataset includes recordings from 8 participants (6 males and 2 females with mean age 30.6 and 28.2 years). The total duration of the recordings is 90 min (6.5 min per protocol). Six participants recorded the protocol twice, and the other two only once; the other two were unavailable at the day of the second recordings. Ground truth labels were assigned based on the time-stamps as well as audio and visual inspection of the captured signals; positive class was assigned on entire eating sessions (e.g. the entire time during which an apple is consumed is marked as positive). Prior probability is 0.45 corresponding to 40 min of eating time and 50 min of non-eating time.

4.1 PALT Evaluation Results

We first perform a baseline k-fold cross validation (CV) experiment on the entire dataset as an estimation of intra-subject effectiveness. Feature vectors

Table 1. Evaluation results of cross-validation (CV) and leave-one-subject-out (LOSO) experiments for baseline and PALT SVM models on the entire dataset.

	Precision	Recall	F1 score	Accuracy	SVs
CV baseline	0.89	0.89	0.89	0.90	33,552
CV PALT@100	0.83	0.89	0.86	0.87	232
CV PALT@800	0.85	0.90	0.87	0.88	1,633
LOSO baseline	0.84	0.81	0.81	0.83	31,152
LOSO PALT@100	0.82	0.79	0.79	0.83	233
LOSO PALT@800	0.81	0.82	0.80	0.83	1,632

are randomly partitioned into k folds. For each fold, an SVM model is trained on the other $k-1$ folds and is used to predict the labels of the fold. We set $k=14$ so that the number of feature vectors of each fold is approximately equal to the length of the experimental protocol. The evaluation is performed per feature vector. Before each training, each feature is linearly normalised to $[0,1]$, and the same transformations are applied to the evaluation set. Precision, recall, F1 score, accuracy, and number of SVs are shown as “CV baseline” of Table 1. To evaluate PALT, we repeat the CV experiment and at each iteration we start with an initial training set $\mathcal{T}_0$ of 40 feature vectors (20 positive and 20 negative). We run PALT for 800 rounds. Rows “CV PALT@m” of Table 1 show the results after $m=100$ and $m=800$ rounds.

To evaluate our approach for inter-subject effectiveness we repeat the experiments in leave-one-subject-out (LOSO) fashion. Similar to CV, the dataset is partitioned to folds where each fold now contains data from a single participant. Thus, 8 partitions are created. Row “LOSO baseline” of Table 1 shows the results, and rows “LOSO PALT@m” show the results after $m=100$ and $m=800$ rounds of PALT. Figure 3a shows the mean (per subject) accuracy for the LOSO PALT experiments across all 800 rounds. LOSO baseline accuracy is also shown for comparison. The thinner lines show ±1 standard deviation.

During the LOSO baseline experiment, we further partitioned the available training data of each iteration randomly, in a 70%–30% ratio for hyper-parameter search. However, we have found that for 7 out of the 8 participants, the optimal values are $C=1$ for the SVM, and $\gamma = 10 \cdot D^{-1}$ for the RBF kernel, where $D=16$ is the number of features. We have thus selected these values for all of our experiments.

Baseline models outperform PALT, however the difference in effectiveness is rather small. For CV, PALT after 800 rounds is only 2 percentage units lower in F1 score and accuracy, while for LOSO, PALT achieves the same accuracy 0.83 and only one percentage unit less in F1 score. However, PALT models include approximately 30 times less SVs compared to baseline.

Table 2. Evaluation results of leave-one-subject-out (LOSO) experiments for IALA, using the baseline and PALT to create the initial model.

	Precision	Recall	F1 score	Accuracy	SVs
LOSO_6 baseline	0.84	0.82	0.81	0.82	25,043
LOSO_6 PALT	0.87	0.66	0.72	0.82	1,633
LOSO_6 baseline + IALA	0.84	0.83	0.82	0.83	25,038
LOSO_6 PALT + IALA	0.88	0.80	0.83	0.85	1,652

4.2 IALA Evaluation Results

To evaluate IALA we perform additional experiments on the six participants that recorded the protocol twice. The prior probability in this subset is 0.46. The LOSO baseline and PALT experiments are repeated on the six participants subset. We then use one of the protocol recordings to simulate the stream of feature vectors, and the second protocol to evaluate effectiveness. Results before and after running IALA are shown in Table 2. IALA improves baseline accuracy from 0.82 to 0.83, and PALT accuracy from 0.82 to 0.85. PALT F1 score is improved from 0.72 to 0.83, however this huge improvement is caused by the low recall (and thus F1) of the PALT models; the PALT + IALA F1 score of 0.83 is higher than the baseline 0.81. These results are quite encouraging given the short duration of the protocol used for simulating streaming mode (roughly (6 min). In addition, inter-subject variance of accuracy (see Fig. 3b) descreases as SVM models are adapted per-subject, indicating robust convergence to more effective models. The highest effectiveness among all LOSO experiments is achieved by the combination of PALT and IALA.

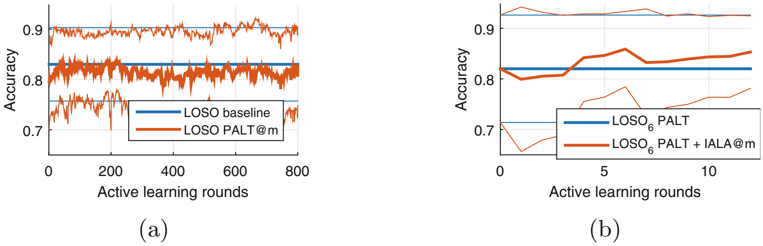


Fig. 3. Accuracy across active learning rounds vs. baseline. (a) LOSO baseline vs. LOSO PALT. (b) LOSO_6 PALT vs. LOSO_6 PALT + IALA.

5 Conclusions

This paper presents a chewing detection system based on audio signals from an off-the-shelf bone-conduction microphone connected to an Android smart-phone.

The system uses AL for two tasks; to create and deploy a classification model with fewer SVs that requires reduced computational resources, and to enable per-user adaptation of the deployed model requiring minimal and real-time user feedback. Validation on an experimental dataset recorded in lab conditions shows inter-subject accuracy of 0.85 using user-adapted models and parsimonious initial SVM models. Future work includes evaluation the proposed system on a larger dataset under free-living conditions.

Acknowledgements. The work leading to these results has received funding from the European Communitys Health, demographic change and well-being Programme under Grant Agreement No. 727688, 01/12/2016 - 30/11/2020.

References

1. Amft, O., Stäger, M., Lukowicz, P., Tröster, G.: Analysis of chewing sounds for dietary monitoring. In: Beigl, M., Intille, S., Rekimoto, J., Tokuda, H. (eds.) UbiComp 2005. LNCS, vol. 3660, pp. 56–72. Springer, Heidelberg (2005). https://doi.org/10.1007/11551201_4
2. Archer, E., Hand, G.A., Blair, S.N.: Validity of us nutritional surveillance: national health and nutrition examination survey caloric energy intake data, 1971–2010. PloS ONE **8**(10), e76632 (2013)
3. Fontana, J.M., Farooq, M., Sazonov, E.: Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior. IEEE Trans. Biomed. Eng. **61**(6), 1772–1779 (2014)
4. Kyritsis, K., Tatli, C.L., Diou, C., Delopoulos, A.: Automated analysis of in meal eating behavior using a commercial wristband IMU sensor. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (July 2017, to appear online)
5. Papapanagiotou, V., Diou, C., Delopoulos, A.: Chewing detection from an in-ear microphone using convolutional neural networks. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (July 2017, to appear online)
6. Papapanagiotou, V., Diou, C., Lingchuan, Z., van den Boer, J., Mars, M., Delopoulos, A.: Fractal nature of chewing sounds. In: Murino, V., Puppo, E., Sona, D., Cristani, M., Sansone, C. (eds.) ICIAP 2015. LNCS, vol. 9281, pp. 401–408. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23222-5_49
7. Papapanagiotou, V., Diou, C., Zhou, L., van den Boer, J., Mars, M., Delopoulos, A.: A novel chewing detection system based on PPG, audio and accelerometry. IEEE J. Biomed. Health Inf. **21**, 607–618 (2016)
8. Sazonov, E., Schuckers, S., Lopez-Meyer, P., Makeyev, O., Sazonova, N., Melanson, E.L., Neuman, M.: Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior. Physiol. Meas. **29**(5), 525 (2008)
9. Sazonov, E.S., Fontana, J.M.: A sensor system for automatic detection of food intake through non-invasive monitoring of chewing. IEEE Sens. J. **12**(5), 1340–1348 (2012)
10. Shuzo, M., Komori, S., Takashima, T., Lopez, G., Tatsuta, S., Yanagimoto, S., Warisawa, S., Delaunay, J.J., Yamada, I.: Wearable eating habit sensing system using internal body sound. J. Adv. Mech. Des. Syst. Manuf. **4**(1), 158–166 (2010)
11. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. **2**, 45–66 (2001)