Chewing detection from an in-ear microphone using convolutional neural networks

Vasileios Papapanagiotou, Christos Diou, Anastasios Delopoulos

Abstract— **Detecting chewing sounds from a microphone placed inside the outer ear for eating behaviour monitoring still remains a challenging task. This is mainly due the difficulty in discriminating non-chewing sounds (e.g. speech or sounds caused by walking) from chews, as well as due to to the high variability of the chewing sounds of different food types. Most approaches rely on detecting distictive structures on the sound wave, or on extracting a set of features and using a classifier to detect chews. In this work, we propose to use feature-learning in the time domain with 1-dimensional convolutional neural networks for for chewing detection. We apply a network of convolutional layers followed by fully connected layers directly on windows of the audio samples to detect chewing activity, and then aggregate individual chews to eating events. Experimental results on a large, semi-free living dataset collected in the context of the SPLENDID project indicate high effectiveness, with an accuracy of** 0.980 **and F1 score of** 0.883**.**

I. Introduction

Automatic detection of eating activity using wearable sensors is a research problem that has been active for the past decade, with several promising applications related to the prevention of obesity [1], or more generally for the adoption of a more healthy lifestyle. Proposed sensors include microphones worn inside the ear [2] or around the neck [3], strain sensors [4], electromyography sensors [5], or wrist-worn motion sensors [6].

Detection of chewing sounds from microphones is one of the earliest explored modalities. Most commonly, audio is captured by a microphone placed inside the ear, usually inside the outer ear canal, as this positioning naturally amplifies body generated sounds such as chewing sounds, and also attenuates external and environmental sounds. Various approaches have been proposed for extracting chewing related information from audio signals. In [7], a total of seven algorithms are evaluated; some of them compute a rectified version of the audio energy signal and then detect peaks as chews, while others use the same rectified energy signal to detect periodical events in the 0 - $2\,Hz$ frequency range. In the works of [3], [8] audio is processed and features are extracted, including spectral, wavelet-based, and higher-order statistics; the features are then used by a classfier to characterize audio segments as swallowing [3] or chewing [8] respectively.

All these are based on "hand-crafted" features extracted from the audio signal and on well-studied classification algorithms, such as Support Vector Machines. However recent advances in machine learning and especially deep neural networks now enable evaluation of feature learning approaches for the detection of eating. Deep learning and convolutional neural networks (CNN) in particular, have been extremely successful in computer vision applications (e.g. [9]). Deep neural networks and CNNs have also been applied to audio signals [10]. In [11] a deep belief network was trained on mel-frequency cepstral coefficients (MFCC) for automatic speech recognition (ASR); these features were chosen to reduce the computational burden (compared to using raw audio segments) since they are known to be discriminative for speech applications. Experimental results on the TIMIT[1] database improved over state-of-the-art approaches based on Hidden Markov Models by achieving the lowest error rate.

In the more recent work of [12], [13] CNNs were applied directly on raw audio signals for ASR. In [12] authors used CNNs combined with conditional random fields to train an end-to-end system, improving over an MFCC-based baseline approach both on TIMIT and Wall Street Journal (WSJ) datasets. In [13] the authors also evaluated the cross-domain effectiveness of the learned features on TIMIT and WSJ. Error rates were similar when testing on TIMIT (approximately 32%); however, testing on WSJ seemed to benefit when features were learned on WSJ (6.7%) instead of TIMIT (10.1%).

In this work we apply CNNs on audio recordings from an ear-worn microphone for the task of chewing detection. Contrary to MFCC in ASR problems, no feature stands out as being discriminative for detecting chews (although several have been tested). We thus apply the CNNs on a lightly pre-processed version of the raw audio signals. The output of the CNN is smoothed and aggregated to chewing bouts and then to eating sessions (based on the aggregation method proposed in [8]). We apply our method to a large (60 hours) semi-free living dataset recorded in the context of the SPLENDID project. Experimental results show that the CNN approach is very promising, since it achieves a significant effectiveness improvement compared to existing state-of-the-art approaches. The rest of this paper is organised as follows. Section II presents the processing pipeline, focusing on the CNN architecture. Section III presents the dataset, evaluation methods and experimental results. Finally, Section IV concludes the paper.

*The work leading to these results has received funding from the European Community's Health, demographic change and well-being Programme under Grant Agreement No. 727688, 01/12/2016 - 30/11/2020 (http://bigoprogram.eu/).

*We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

The authors are with the Multimedia Understanding Group, Information Processing Laboratory, Aristotle University of Thessaloniki, Greece

[1] `https://catalog.ldc.upenn.edu/ldc93s1`

II. Designing a CNN for chewing detection

Initially, we perform a pre-processing step to the audio signals. The original sampling frequency of $48\,kHz$ is too high and would significantly increase the computational requirements due to longer filters and increased number of neurons. We thus apply a low-pass filter on the signals and downsample them at $2\,kHz$; this frequency is also used in [8], [14] (which we use as reference in our experiments). Furthermore, we apply an FIR high-pass filter with cut-off frequency at $20\,Hz$ to remove the very low frequencies of the signal caused by amplifier drift and environmental conditions (e.g. blowing wind while the participants were walking outdoors).

The input to the CNNs are audio segments (windows) of the filtered audio signal. We experiment with different window lengths based on two different approaches: (a) longer windows, e.g. $5\,sec$, that capture the periodic/rhythmic nature of consecutive chews, and (b) shorter windows, e.g. $1\,sec$, that capture the morphology of a single chew. The window step is fixed to $0.1\,sec$. Experimenting with various configurations has shown that five convolutional layers are sufficient to yield promising results without excessively increasing the number of model parameters.

The number of filters of the convolutional layers increases exponentially with depth (except for the fifth layer); in particular we set 8, 16, 32, 64 and 64 filters for each layer. The filters' length is set to 16 for layers 1-4 and is approximately doubled for the fifth, to account for the duration of one chew (approximately $0.5\,sec$. At each layer, full convolution is performed and Rectified Linear Unit (ReLU) activations are used. Performing full convolution on the last layer enables the detection of chews in the window without having to exactly align it to the filters; this allows for the relatively long window step of $0.1\,sec$, which we use.

Each convolutional layer is followed by a max pooling layer with a step of either 2 or 4. These values are chosen differently for each window length; for the longest window ($5\,sec$) the step is set to 4 for all layers, whereas for the shortest window only the fifth layer's step is set to 4. The different configurations are also shown in Table I.

Each configuration includes three fully connected layers after the fifth max pooling layer; the corresponding numbers of neurons are 200, 200 and 2. The two outputs p_1 and p_2 of the last softmax layer correspond to chewing and non-chewing respectively. We also apply dropout [15] on the two layers of 200 neurons to avoid overfitting. The dropout probability is set to 0.5. The training step maximises the cross entropy of the of the CNN output and the window label for each training batch

$$H = -\sum_{i=1}^{n} c(i) * \log(p_1(i)) + (1 - c(i)) * \log(p_2(i)) \quad (1)$$

where n is the batch length (we have set $n = 16$). The indicator $c(i)$ is the ground truth label for the i-th window of the training batch based on whether the window's centre point (in time) lies inside an eating event ($c(i) = 1$) or not ($c(i) = 0$). The windows of each batch are chosen randomly, however we make sure that each batch contains 50% positive windows. In the experiments, we use TensorFlow[2] and the Adam optimiser [16] with an initial step of 16^{-4} and train for $5 \cdot 10^5$ epochs.

TABLE I: CNN architectures per window length; notation is "[no. of filters] × [filter length], [max pooling step]".

	5 sec	**3 sec**	**2 sec**	**1 sec**
1	8×16, 4	8×16, 2	8×16, 2	8×16, 2
2	16×16, 4	16×16, 2	16×16, 2	16×16, 2
3	32×16, 4	32×16, 4	32×16, 2	32×16, 2
4	64×16, 4	64×16, 4	64×16, 2	64×16, 2
5	64×39, 4	64×23, 4	64×31, 2	64×31, 4
6	200 fully connected neurons with 0.5 dropout probability			
7	200 fully connected neurons with 0.5 dropout probability			
8	2 fully connected neurons (outputs y_1, y_2) and softmax (p_1, p_2)			

In order to detect eating events, we apply the trained model on the test audio signals and derive the output sequence for the chewing class $y_1(n)$ (output before softmax), at the rate of $10\,Hz$ based on the selected window step. We apply smoothing using a $3\,sec$ filter from a normalised hamming window (we have experimented with lengths from 0-10 sec) to produce a smoothed output $y_1'(n)$ and then threshold the $y_1'(n)$ sequence with 0. We interpret positive values as chewing and negative values as non-chewing. This boolean indicator signal is then aggregated into chewing bouts using the method of [8].

More specifically, detections which are closer than $2\,sec$ are merged and the ones with duration less than $5\,sec$ are discarded. This aggregates detections into chewing bouts. Then, chewing bouts are aggregated into snacks by merging successive chewing bouts which are up to $1\,min$ apart. A final pruning step removes snacks which last less than $30\,sec$ or are covered by chewing bouts less than 25%. A MATLAB implementation of the aggregation method is available through GitHub[3].

III. Dataset & experimental evaluation

Audio signals were collected at Wageningen University during the summer of 2015. The recordings were performed in the context of the EU funded SPLENDID project, using an in-ear microphone, Knowles FG-23329-D65 housed an a commercial ear bud, at $48\,kHz$. The ear bud also houses a (photoplethysmography) PPG sensor, and a belt-mounted 3D accelerometer was worn by the participants. Both the PPG and the accelerometer signals were recorded at $\frac{64}{3}\,Hz$. The dataset includes 26 recording sessions from 14 participants, with total duration approximately 60 hours. During the recording trials, each participant had two main meals; they started with lunch, followed by a unscripted period, and conluded with dinner at the end of the day. During the unscripted period they were free to live the university premises, but were instructed to perform some

[2] `tensorflow.org`

[3] `github.com/mug-auth/chewing-detection-challenge`

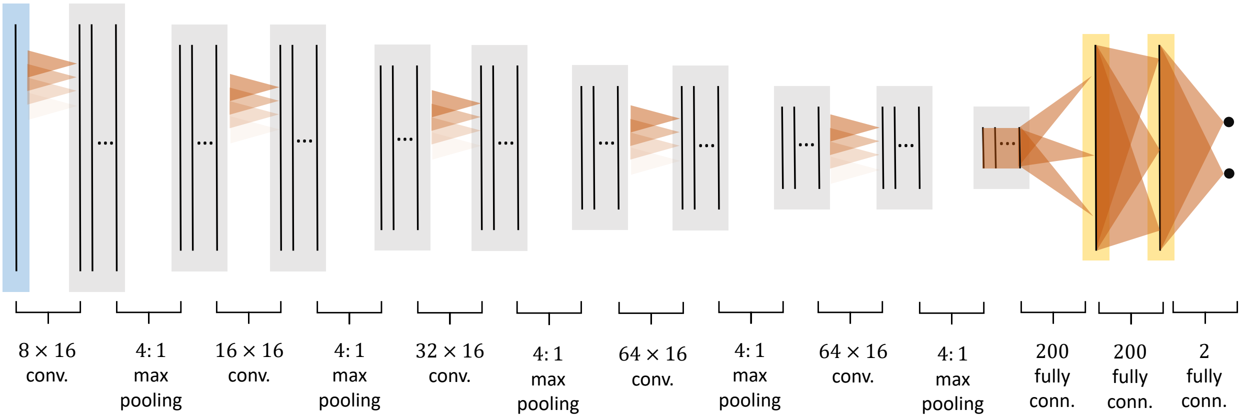

Fig. 1: The proposed CNN architecture for the $5\,sec$ input window.

physical activities and have at least three snacks. More details about the sensors and the dataset can be found in [17]. We selected 4 participants, each with a single recording session, as the test set; the recording sessions from the remaining 10 participants were used as the training set. This selection was made randomly, prior to any experimentation and was not changed thereafter.

We evaluate the performance of our method using five metrics: precision, recall, accuracy, weighted accuracy and F1 score. We use weighted accuracy with weight $w = 6.9$ based on the dataset's prior probability for the chewing class. We present three evaluation methods, as per [8]:

- Average duration-based evaluation: where the metrics are computed for each of the four participants of the test set and then averaged
- Cumulative duration-based evaluation: where the metrics are computed once for the entire duration of the dataset
- Event-based evaluation: where a one-to-one matching is performed based on overlapping duration (only one match per eating event, and only if the overlap exceeds 75%.

Results are shown on Table II, sorted by decreasing F1 score. We also include results from [8], for comparison purposes. "Audio" refers to an algorithm which uses only the audio signal, while the "Fusion+" algorithm (included for reference) uses also PPG and accelerometer signals. The training and test is performed on the same split we use for the CNNs. In addition, we also include "leave-one-subject-out" results for these two algorithms, denoted "Audio (LOSO)" and "Fusion+ (LOSO)" respectively. Thus, only "Audio" is directly comparable with the proposed method (since they use the same input).

Results show that the CNN with $5\,sec$ input window achieves 0.89 precision and 0.92 recall, and 0.95 weighted accuracy. These results are better than all other CNNs as well as the previous "Audio" approach of [8]. Furthermore, the proposed CNN approach outperforms "Fusion+" for most metrics of duration-based evaluation methods. This is surprising, given that the "Fusion+" approach uses additional sources of information (PPG and accelerometer, in addition to audio).

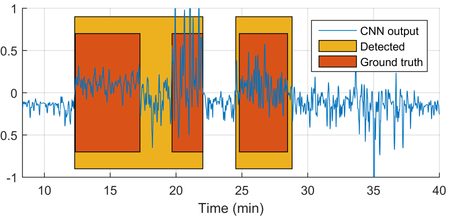

Fig. 2: An example of the output of the $5\,sec$ input window CNN, and the derived detections of eating events. The first detected event is not matched to any of the first two ground truch events due to low overlapping duration.

From the other CNNs, the $2\,sec$ input window configuration also demonstrates high effectiveness both in precision and recall. It is noteworthy that the 5 and $2\,sec$ CNNs seem to be complementing each other in recall and precision respectively; this could indicate that a suitable combination of the two configurations could lead to very high effectiveness in both metrics combined, however we did not test mixing different input window lengths in this work.

For the event-based detection, the fusion of PPG, audio and accelerometer signals of [8] still yields the highest F1 score of 0.734. The $5\,sec$ CNN effectiveness is comparable however, achieving 0.706 F1 score; this difference is caused mainly by the relatively lower recall of 0.75. This in turn is mainly due to certain long eating events detected by the CNN as multiple, fragmented eating events, yielding false positives in the event-based evaluation. Such an example is shown in Figure 2, where the first detected eating event is not matched to any of the first two ground truth events, yielding one false positive and two false negative events (the second detected event is matched with the third ground truth event yielding one true positive).

TABLE II: Evaluation results for the proposed CNN architectures for precision, recall, accuracy, weighted accuracy and F1 score. including results from [8] for audio only (Audio) and combined audio, PPG and accelerometer (Fusion+).

(a) Average duration-based evaluation

	prec.	rec.	acc.	w. acc.	F1
$5\,sec$	0.796	0.991	0.980	**0.984**	**0.883**
$3\,sec$	0.576	**0.999**	0.945	0.961	0.731
$2\,sec$	**0.991**	0.793	**0.984**	0.925	0.881
$1\,sec$	0.988	0.692	0.976	0.889	0.814
Audio	0.215	0.700	0.708	0.704	0.329
Audio (LOSO)	0.633	0.809	0.880	0.861	0.650
Fusion+	0.226	0.687	0.729	0.713	0.340
Fusion+ (LOSO)	0.794	0.807	0.938	0.892	0.761

(b) Cumulative duration-based evaluation

	prec.	rec.	acc.	w. acc.	F1
$5\,sec$	0.890	**0.927**	**0.976**	**0.955**	**0.908**
$3\,sec$	0.812	0.883	0.959	0.927	0.846
$2\,sec$	0.938	0.811	0.969	0.902	0.870
$1\,sec$	**0.956**	0.745	0.963	0.870	0.838
Audio	0.294	0.714	0.263	0.572	0.417
Audio (LOSO)	0.476	0.811	0.861	0.840	0.600
Fusion+	0.321	0.643	0.273	0.537	0.429
Fusion+ (LOSO)	0.702	0.800	0.931	0.875	0.748

(c) Event-based evaluation

	prec.	rec.	F1
$5\,sec$	0.667	0.750	0.706
$3\,sec$	0.348	0.500	0.410
$2\,sec$	0.429	0.563	0.486
$1\,sec$	0.600	0.562	0.581
Audio	0.215	0.700	0.329
Audio	0.447	**0.837**	0.583
Fusion+	0.408	0.549	0.468
Fusion+ (LOSO)	**0.677**	0.802	**0.734**

IV. Conclusions & future work

In this work we have presented a method for detection chewing activity from an in-ear microphone using CNN. To the best of our knowledge, this is the first attempt of using CNN for chewing detection. We experiment with various configurations and propose a network of five convolutional layers followed by two fully connected layers and two outputs. We train and evaluate our method on a large, challenging, and semi-free living dataset collected in the context of SPLENDID project, and obtain high effectiveness results, achieving 0.984 weighted accuracy and 0.883 F1 score for average duration-based evaluation (0.955 and 0.908 for cumulative duration-based evaluation). The best CNN configuration of 5 sec input window outperforms the current state-of-the-art approach for the same audio signals, and also outperforms, in most cases, a multi-sensor approach that combines the audio with PPG and accelerometer signals. Future work includes exploring the concurrent detection of chewing from multiple windows as well as the application of CNNs for the fusion of different sources (PPG and accelerometer) for chewing detection.

References

[1] C. Maramis, C. Diou, I. Ioakeimidis, I. Lekka, G. Dudnik, M. Mars, N. Maglaveras, C. Bergh, and A. Delopoulos, "Preventing obesity and eating disorders through behavioural modifications: the splendid vision," in *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*. IEEE, 2014, pp. 7–10.

[2] O. Amft, M. Kusserow, and G. Trster, "Bite weight prediction from acoustic recognition of chewing," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 6, pp. 1663–1672, June 2009.

[3] E. S. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. L. Melanson, and M. R. Neuman, "Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 626–633, March 2010.

[4] M. Farooq and E. Sazonov, "Comparative testing of piezoelectric and printed strain sensors in characterization of chewing," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2015, pp. 7538–7541.

[5] O. Amft and G. Troster, "Methods for detection and classification of normal swallowing from muscle activation and sound," in *2006 Pervasive Health Conference and Workshops*, Nov 2006, pp. 1–10.

[6] M. Mirtchouk, C. Merck, and S. Kleinberg, "Automated estimation of food type and amount consumed from body-worn audio and motion sensors," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '16. New York, NY, USA: ACM, 2016, pp. 451–462. [Online]. Available: http://doi.acm.org/10.1145/2971648.2971677

[7] S. Päßler and W.-J. Fischer, "Evaluation of algorithms for chew event detection," in *Proceedings of the 7th International Conference on Body Area Networks*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2012, pp. 20–26.

[8] V. Papapanagiotou, C. Diou, L. Zhou, J. van den Boer, M. Mars, and A. Delopoulos, "A novel chewing detection system based on ppg, audio and accelerometry," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, no. 99, pp. 1–1, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[10] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[11] A. r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan 2012.

[12] D. Palaz, R. Collobert, and M. Magimai-Doss, "End-to-end phoneme sequence recognition using convolutional neural networks," *CoRR*, vol. abs/1312.2137, 2013. [Online]. Available: http://arxiv.org/abs/1312.2137

[13] D. Palaz, M. M. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4295–4299.

[14] V. Papapanagiotou, C. Diou, Z. Lingchuan, J. van den Boer, M. Mars, and A. Delopoulos, "Fractal nature of chewing sounds," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, vol. 9281, pp. 401–408. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-23222-5_49

[15] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[17] V. Papapanagiotou, C. Diou, and A. Delopoulos, "The splendid chewing detection challenge," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul 2017.