Eating Sound Dataset for 20 Food Types and Sound Classification Using Convolutional Neural Networks

Jeannette Shijie Ma
LIACS, Leiden University
Leiden, Netherlands
s.ma.3@umail.leidenuniv.nl

Marcello A. Gómez-Maureira
LIACS, Leiden University
Leiden, Netherlands
m.a.gomez.maureira@liacs.leidenuniv.nl

Jan N. van Rijn
LIACS, Leiden University
Leiden, Netherlands
j.n.van.rijn@liacs.leidenuniv.nl

ABSTRACT

Food identification technology potentially benefits both food and media industries, and can enrich human-computer interaction. We assembled a food classification dataset consisting of 11,141 clips, based on YouTube videos of 20 food types. This dataset is freely available on Kaggle. We suggest the grouped holdout evaluation protocol as evaluation method to assess model performance. As a first approach, we applied Convolutional Neural Networks on this dataset. When applying an evaluation protocol based on grouped holdout, the model obtained an accuracy of 18.5%, whereas when applying an evaluation protocol based on uniform holdout, the model obtained an accuracy of 37.58%. When approaching this as a binary classification task, the model performed well for most pairs. In both settings, the method clearly outperformed reasonable baselines. We found that besides texture properties, eating action differences are important consideration for data driven eating sound researches. Protocols based on biting sound are limited to textural classification and less heuristic while assembling food differences.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Information systems** → **Clustering and classification**; • **Applied computing** → *Sound and music computing*.

KEYWORDS

sound classification, food classification, neural networks, eating sound, sound dataset

ACM Reference Format:
Jeannette Shijie Ma, Marcello A. Gómez-Maureira, and Jan N. van Rijn. 2020. Eating Sound Dataset for 20 Food Types and Sound Classification Using Convolutional Neural Networks. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3395035.3425656

1 INTRODUCTION

Food is one of the most important elements that directly interact with our body. Humans have evolved with delicate perception of food in order to survive and thrive. The sound of food is tightly connected to the textural perception of food, and provides important information on food quality (such as freshness, water content, and palatability [5, 14, 16]). Eating sounds have also been studied for their rich application potential [4, 15]. Diet tracking, for example, is an area that could benefit greatly from classifying food based on sound. Tracking of food can be important to monitor personal health in daily life as well as in hospital settings to inform about the nutritional excess or insufficiency of diets. At present, diet tracking tends to rely on manually entering information for each meal. Researchers within the nutrition and eating behaviour fields have been trying to develop a more automated way to detect diet and eating behaviours [15]. Dacremont (1995) looked into sound spectrum features of 8 different foods eating by 60 subjects [3]. In another research, Shuzo *et al.* (2010) successfully applied a sound classification method for a portable eating behaviour detector with a bone-conduction microphone [13]. However, the sound samples in these studies were recorded in a carefully controlled situations with high recording quality. The results might only be applicable on body-contacting detectors. To the best of our knowledge, there is no large-scale benchmark on eating sounds which resemble our daily eating situations.

More recently, people have started to record eating sounds as part of 'ASMR' (autonomous sensory meridian response) videos in an effort to cause a pleasant tingling sensation in a listener. Setting aside the fact that the act of eating food is necessarily creating sound, the sound of eating can in itself be considered a form of communication that provides information. In this case, the sound of eating can provide information about what is being eaten. This 'communication' is not only available to human listeners, but can also be captured and classified by computers.

Convolutional neural networks [6] have been applied to classify large-scale featured noise like urban environment sounds [11, 12]. In these studies, large manually labelled sound data-sets were used to train the model in classifying different sound sources. These classification experiments often achieve excellent performances since the sound categories have significantly different features, which are also easy to distinguish for human listeners.

Our research aims to evaluate the performance of convolutional neural networks on food eating sound classification with online public-sourced training data, representing various eating conditions, behaviours and recording qualities. Our contributions are the following:

(1) We assemble a public sourced sound dataset from different food types.
(2) We propose a corresponding evaluation protocol, based on grouped holdout evaluation.

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425656

(3) We experiment with convolutional neural networks to assess baseline performance.
(4) We analyse distances of various food types using clustering methods.

In the following sections we will review related work and explain our approach in detail.

2 RELATED WORK ON EATING SOUND CLASSIFICATION

In previous work, sound features like amplitude, number of sound bursts and mean peak height were evaluated to characterise the texture of food products [4]. Aside of time-based parameters, spectrum composition of eating sounds were also studied in order to understand the distinct sound features generated with certain food textures [3]. The sound and texture correlations calculated from these studies focused on preciseness and general representation. However, as eating sound is generated through complex movements and various mouth structures, new studies require a more generalised view point to observe the textural sound features.

With the potential application of hospital eating behaviour monitoring, more recent studies focused on the development of wearable eating monitors [15]. Based on previous studies on the relation of sound parameters and textural perceptions, eating monitor research evolved from feature based towards more data driven methods [1, 13]. These studies used high recording quality, involving bone-conducted microphone or controlled experiment cabins. Recent research explored gathering data in daily situations instead of laboratories, while still using limited numbers of participants and high quality recording methods [8, 10].

To the best of our knowledge, there have not been any top-down data driven studies for generalised eating classification. Because of challenges in recruiting participants and monitoring them, the amount of data collected in previous studies is often limited.

3 DATA COLLECTION AND PREPARATION

All the sound clips used in this project were obtained from public videos. The subsections below explain the protocol of video and clip collection as well as the file construction of the dataset.

3.1 Video Collection and Clip Selection

The video materials were collected from YouTube, relying on its availability and amount of content generated by the eating-themed channels. Twenty food categories were selected from the top search results of the term 'eating sound' based on their popularity and food types. This criterion kept a balance of food types and made sure that there are sufficient videos available for each food type. By searching with each food type with 'eating sound' (e.g. 'aloe eating sound'), 11 to 14 videos of each of these 20 food types were downloaded in their highest quality available (resulting in total 246 videos). The videos were screened to make sure that the content is aligned with the title. All the videos were recorded inside a room, but with various space properties (room reverb, obstruction etc.), food varieties (e.g. burgers with/without salad), recording quality and eating behaviours.

For each video, all available eating sounds were located and processed into clips by cutting out talking, cutlery and packaging sounds. Long clips with a repetitive sound profile were separated into smaller pieces of similar lengths. After that, peak normalisation gain targeting -1db was applied to all the clip regions (where 0 db represents the distortion edge). Each food category yielded 279 to 873 clips, adding up to 11,141 clips in total, ranging from 1 to 22 seconds per clip.

The food types are listed below (with the number of clips indicated in parentheses). Each food type involves a range between 11 to 14 source videos that were used to create the clips: Aloe (547), Burger (596), Cabbage (500), Candied fruits (807), Carrots (661), Chips (720), Chocolate (291), Drinks (293), Fries (645), Grapes (580), Gummies (679), Ice-cream (728), Jelly (443), Noodles (412), Pickles (873), Pizza (610), Ribs (489), Salmon (502), Soup (279), Chicken wings (505). In order to make full use of the assembled clips, we did not balance the dataset. Pickles is the largest class, representing roughly 7.8% of the clips. Chocolate is the smallest class, representing roughly 2.6% of the data.

3.2 File Construction

The selected and labelled clips were published on kaggle.com under PDDL license for public experiments.[1] All the clips are in the PCM WAV format, using a sample rate of 44.1kHZ and 24 bit depth. The dataset consist of 20 folders, each containing all the clips of that food. The clips were named with the food name, followed by the video source and then clip number (e.g. *aloe*_10_02.*wav* is the 2nd clip from the 10th video of aloe). The data can be pre-processed in different ways and used for various research or creation purposes.

4 EXPERIMENTS AND RESULTS

Our study used the aforementioned dataset to experiment with two neural networks training tasks:

(1) 20-way classification task: trained by all data from 20 food types. Given a sound clip, the model need to identify which food type is the sound source. A majority class classifier would obtain an accuracy of 7.8%.
(2) pairwise classification task: Performed for each pairs of the 20 categories (in total 190 pairs). Trained by one pair at a time (e.g., aloe vs. burger). Given a sound clip, the model need to tell which of the two food types it is. We would expect the the majority class classifier to obtain an accuracy between roughly 50% (for balanced pairs) and 75% (for the least balanced pair, i.e., pickles vs. chocolate).

This section explains the process of data preparation, protocols, model training of each task and their corresponding evaluation results.

4.1 Data Pre-Processing

We translated the clips into a mel-frequency spectrogram using the Python LibROSA module [9]. As such, each sound clip is represented as an image. We further used the image data pre-processing functions of Keras [2] to get the spectrogram data ready for model training. This method was adapted from previous research of large-scale noise classification with various sample lengths [7].

[1] Eating Sound Collection (Version 1), Available on: https://www.kaggle.com/mashijie/eating-sound-collection

4.2 Model Construction

We built a sequential neural network model with the ADAM optimiser, as implemented in Keras [2]. The network architecture was loosely inspired by the research on convolutional neural networks for large-scale audio classification [7]. The network was made up with six convolutional layers which have increasing filter density. Dropout and pooling layers were included to compensate overfitting and improve model efficiency. The model was trained with a learning rate of 0.0005 and applied categorical cross entropy as loss function. After evaluating the trial results, the rate for each dropout layer was tuned from 0.5 to 0.6 for better performance.

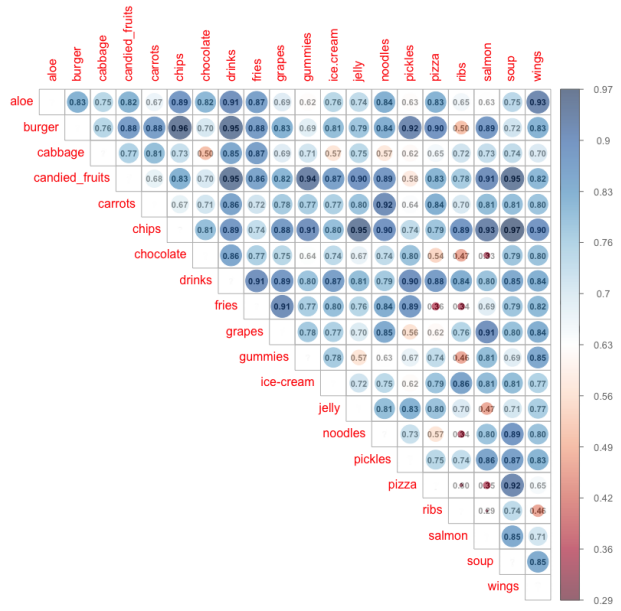

Figure 1: Pairwise classification result (accuracy), using the grouped holdout protocol. The numbers in each cell denotes the accuracy of the pairwise classification task on the intersecting food type column and row. The scale on the right shows the color coding of the accuracy values.

4.3 Evaluation Protocols

Each video is split up into a range of clips that are taken from it. Therefore, the protocols of training and validation splitting are different depending on whether the video grouping were considered as affecting element. For this project we compared **uniform** holdout and **grouped** holdout evaluation protocol. For the uniform holdout protocol, clips from various videos are spread uniformly across the training and validation data. As such, the model might pick up patterns related to the person eating the food, rather than the food itself. To alleviate this problem, we also applied the grouped holdout protocol. The training and validation data was split by groups of videos. For each food type, clips from 70 percent of the videos were used for training and the rest were used for validation. This protocol avoided clips from the same video being in both training and testing sets. Therefore, if the model picked up patterns of certain videos (e.g., eating behaviour of the subject), those patterns do not contribute to improve the test results. A similar procedure is used for common benchmarks, such as the MNIST dataset.

4.4 20-way Classification Task

In this task, we compared the performance of the model using the uniform holdout and grouped holdout evaluation protocols. We evaluated on both models using 10 times repetition, to get a stable performance estimate. Table 1 shows the results. Using the uniform protocol, the model achieves 37.58% average accuracy while when using the more challenging grouped protocol, the model achieves 18.5% average accuracy. This is only a first baseline result to validate that there is a learnable concept in the data. We verified that the performance is higher than Majority Class Classifier (7.8%, calculated analytically). Furthermore, it can be seen that the model benefits greatly from having access to different clips from the same video. As such, the more challenging grouped evaluation protocol is important to assess the real model in realistic settings.

Protocol	Majority Class	Convolutional NN
Uniform Holdout	7.8%	37.58% ± 0.7%
Grouped Holdout	7.8%	18.5% ± 1.3%

Table 1: Accuracy of the CNN using both the uniform and grouped evaluation protocol. For comparison, the expected performance of the Majority Class Classifier is also noted.

4.5 Pairwise Classification Task

In this task we focus on the more challenging grouped holdout evaluation protocol. The training reached convergence within 80 epochs. Again, we report the average accuracy of 10 repetitions. Figure 1 shows the results of each pairwise classification task. The average accuracy and standard deviation per food type is shown in Figure 2. Candied fruits, drinks, chip and soup sounds seem to be relatively distinct and can easily be distinguished. On the other hand, chocolate, ribs and salmon sounds seem to be more ambiguous and generally sound more alike. However, the unbalanced nature of the various problems might be a confounding factor. Figure 3 shows the dendrogram of the matrix result. Using accuracy as distance, this graph clustered similarly-sounding food types (difficult to classify). Some food types with similar texture properties are not clustered together as assumed. This result is not aligned with the result of previous work [1] which was mostly aligned with food textural differences.

5 DISCUSSION

Food identification based on sound patterns is a challenging task. Convolutional Neural Networks score on average 18.1% in the 20-way classification task. The pairwise classification tasks achieved various scores where some pairs could be classified up to 97%. Some of the food pairs from the dataset were especially difficult to classify, which may have caused the low performance of the 20-way classification task. Also, the clip separation methodology used in this paper was aimed to avoid unwanted noise, but might

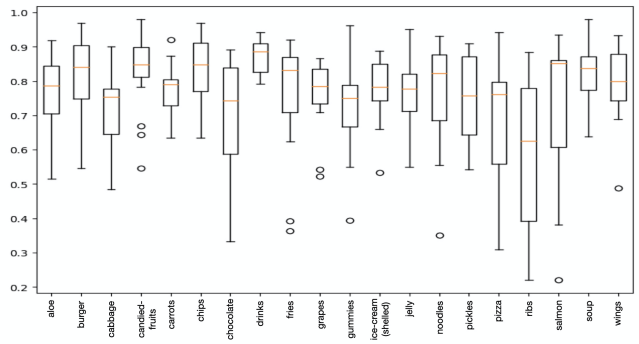

Figure 2: Boxplots of the average pairwise classification performances per food. Each boxplot represents the accuracy of a certain food type compared to all other food types.

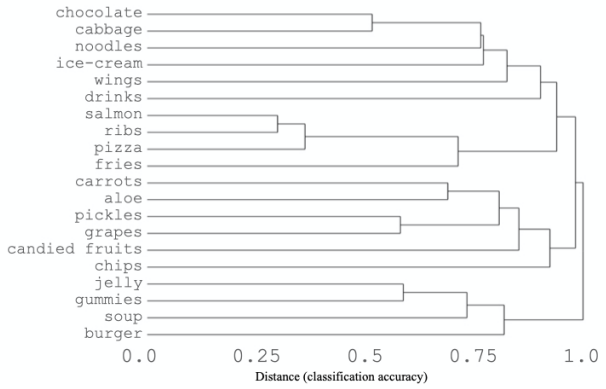

Figure 3: Dendrogram generated from the result matrix using accuracy as distance.

have lost important feature clues in some clips compared to the others. Experiments with longer clips might be an interesting option to explore whether the model learns better with more featured clues while trading off the effect of more noise. The dendrogram result (Figure 3) suggests there might be other audio clues featuring different food types rather than solely texture differences.

6 CONCLUSION

This research evaluated the performance of convolutional neural networks on food eating sound classification, based on online public-sourced training data, representing various real-life eating conditions, behaviours and recording qualities. As part of this study, an eating sound dataset of 20 different food types was collected, curated, and published on Kaggle. The experiment covered both 20-way classification and pairwise classification tasks. When using the grouped holdout evaluation protocol, the neural network could identify certain food types from the 20 categories with 18.5% accuracy. With the uniform protocol, the model achieved 37.58% accuracy, indicating that the model might have learned patterns related to specific videos for the food identification task. As such, we recommend using the grouped holdout evaluation protocol for this dataset. The model achieved promising binary classification performance for many food pairs. However, we note that the fact that the various pairs are imbalanced is a confounding factor. The cluster results of food types show separation of different textural composition for most of the food types. A few pairs of food types with similar texture but different eating actions were distinctly separated. Therefore, aside of food textural differences, more elements of eating behaviour should be considered when studying eating sounds. The existing experimental protocols focusing on biting sounds might eliminate important sound cues present in real-world scenarios. Their inclusion might lead to better sound classification results for the purposes of classifying food types on the basis of sound in an uncontrolled environment.

REFERENCES

[1] Oliver Amft. 2010. A wearable earpad sensor for chewing monitoring. In *Proceedings of IEEE Sensors Conference*. IEEE, 222–227.

[2] François Chollet et al. 2015. Keras. https://keras.io.

[3] C Dacremont. 1995. Spectral composition of eating sounds generated by crispy, crunchy and crackly foods. *Journal of texture studies* 26, 1 (1995), 27–43.

[4] Lisa Duizer. 2001. A review of acoustic research for studying the sensory perception of crisp, crunchy and crackly textures. *Trends in food science & technology* 12, 1 (2001), 17–24.

[5] Ryan S Elder and Gina S Mohr. 2016. The crunch effect: Food sound salience as a consumption monitoring cue. *Food quality and Preference* 51 (2016), 39–46.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

[7] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

[8] Konstantinos Kyritsis, Christos Diou, and Anastasios Delopoulos. 2020. A Data Driven End-to-end Approach for In-the-wild Monitoring of Eating Behavior Using Smartwatches. *IEEE Journal of Biomedical and Health Informatics* (2020).

[9] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference (SciPy 2015)*. 18–25.

[10] Mark Mirtchouk, Dana L. McGuire, Andrea L. Deierlein, and Samantha Kleinberg. 2019. Automated Estimation of Food Type from Body-worn Audio and Motion Sensors in Free-Living Environments. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, Vol. 106. PMLR, 641–662.

[11] Karol J Piczak. 2015. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

[12] Justin Salamon and Juan Pablo Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* 24, 3 (2017), 279–283.

[13] Masaki Shuzo, Shintaro Komori, Tomoko Takashima, Guillaume Lopez, Seiji Tatsuta, Shintaro Yanagimoto, Shin'ichi Warisawa, Jean-Jacques Delaunay, and Ichiro Yamada. 2010. Wearable eating habit sensing system using internal body sound. *Journal of Advanced Mechanical Design, Systems, and Manufacturing* 4, 1 (2010), 158–166.

[14] Zata Vickers. 1991. Sound perception and food quality. *Journal of Food Quality* 14, 1 (1991), 87–96.

[15] Tri Vu, Feng Lin, Nabil Alshurafa, and Wenyao Xu. 2017. Wearable food intake monitoring technologies: A comprehensive review. *Computers* 6, 1 (2017), 4.

[16] Massimiliano Zampini and Charles Spence. 2004. The role of auditory cues in modulating the perceived crispness and staleness of potato chips. *Journal of Sensory Studies* 19, 5 (2004), 347–363.