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

Jeannette Shijie Ma
MSc. Thesis

Media Technology, LIACS, Leiden University
Primary Supervisor: Jan N. van Rijn
Secondary Supervisor: Marcello A. Gómez-Maureira
August, 2020

ABSTRACT

Food identification technology potentially benefits both food and media industries, and improves human-computer interaction. We assembled a food classification dataset based on 246 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. We did trials with partial of dataset to get a sectional view of the data. After that, both 20-way classificaiton and pairwise classification tasks were performed with the aid of cluster compuing environment. 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. We recommend a generation method for further research to better understand the computer's interpretation of eating sound for different food types.

1. INTRODUCTION

Food is one of the most important elements that directly interact with our body. As human kind, we evolved with delicate perception of food, in order to survive and thrive. The sound of food is tightly related to the textural perception of food, and provides important information on food quality (freshness, water content, palatability [5, 14, 16]). Eating sounds have been widely studied by researchers 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) applied a successful sound classification method to apply in a portable eating behaviour detector with a bone-conduction microphone [13]. However, the sound samples in these studies were recorded in carefully controlled situation 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 such sounds as part of 'ASMR' (autonomous sensory meridian response) videos in an effort to cause a pleasant tingling sensation in listeners that enjoy it. 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 also possible to be captured and classified by computers, as illustrated in Figure 1.

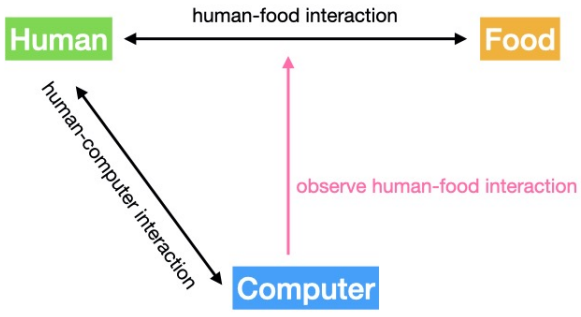

Figure 1. Relationship illustration of computer observing human-food interaction

Related to the task of audio signal processing, convolutional neural networks [6] have been applied to classify large-scale featured noise like urban environment sounds [12, 11]. In these

works, massive manually labelled sound data were used to train the model in classifying different sound sources (e.g. bird sound, traffic sound). These classification experiments usually achieve excellent performances since the sound categories have significantly different features, which are also easy for human to distinguish. Compared to these noises, eating sound of different food can be much more alike and more difficult to classify.

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.

3. We experiment with convolutional neural networks to assess baseline performance.

4. We analyse distances of various food types using clustering methods.

The objectives of this project potentially benefit both food nutrition and media areas. As we found there are trainable elements in the eating sound classificaiton task using massive public data, the classification performance has potential to be improved. With the classification ability, social robots in both online and physical forms will obtain higher human-like empathy while 'listening' to the environment. Thus, related applications on hospital, restaurants, household and retailing situations could also benefit from this 'monitoring' ability. 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]. Besides 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 data in these works were collected in strictly controled environment with specific biting protocols. The sound and texture correlations calculated from these studies aimed at preciseness and general representation. However, as eating sound is generated with complicated movements and various mouth structures, the researchers needed a more generalised view point to observe the textural sound features.

With the potential application of hospital eating behaviour monitoring, more recent researches focused on the development of wearable eating monitors [15]. The objectives of this work is developing an earphone-like portable device which record and analyse the eating behaviour of the users. Based on the previous researches on the relation of sound parameters and textural perceptions, this type of eating monitor researches evolved from feature based towards more data-driven methods [13, 1]. However, these studies still required high recording quality in controlled environments, involving bone-conducted microphone or controlled experiment cabins. On the contrary, recent researches explored gathering data in daily situations instead of laboratories. These works allowed participants recording eating sound in more natural environment, while still using limited numbers of participants and high quality recording methods [10, 8]. Also, the previous data-driven approaches focused on eating behaviour detection instead of food type classification.

To the knowledge of authors, there have not been any auditory-based food type classification studies involving data driven top-down generalised eating sound situations. Because of the 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 data used in this project were taken from public video sources. The subsections below explained the protocol details of video and clip collection as well as the file construction of the dataset.

3.1 Video Collection

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 type of food. By searching with each food name 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. The videos were screened to make sure the contents are aligned with their titles (resulting in total 246 videos). All these 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. Since the style of videos are different, some videos include large amount of talking sound while others only contain eating related sounds. The ambiant noises levels are also different among all the videos. Some videos includes distant traffic sound or air-conditioning sound.

3.2 Clip Selection

For each video, all available eating sounds were located and processed into clips by cutting out talking, cutlery and packaging sounds. In order to better represent food features in the dataset, 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). The normalization standardized the sound power range of all clips, so that the features were evenly spreaded on the later generated spectrograms. Each food category yielded 279 to 873 clips, adding up to 11141 clips in total, ranging from 1 to 22 seconds per clip.

The food types were listed below (with the number of clips indicated in parentheses). Each food type involved 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.

3.3 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 a main folder including 20 subfolders, each containing all the clips of that food type. 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]. Mel-frequency spectrogram ploted three features of sound waves: as shown in Figure2, the horizontal axis represented time in second. The vertical axis represented frequency and color scaled by the power of sound in dB. As such, each sound clip is represented as an image. The same procedures were applied in previous works on noise classification. 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 research with various sample lengths [7].

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

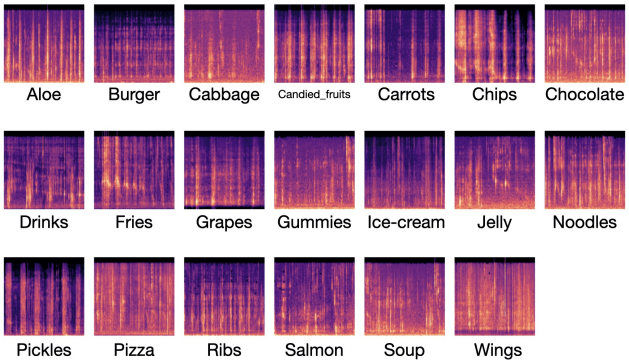

Figure 2. Examples of mel-spectrograms for the 20 foods, displaying time(s) as horizontal axis, frequency(kHz) as vertical axis and power(dB) as color scale.

4.2 Model Construction

We build 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 over-fitting 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.

4.3 Setting 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 introduced the grouped holdout protocol. The training and validation data was split by groups of videos. For each food type, clips from 70 percent 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

We did three experiments for the 20-way classificaiton task: First we compared the uniform and grouped holdout protocol. After that, we compared the full dataset with a reduced dataset using the grouped holdout protocol. This section explained the details and results of these experiments.

Compare uniform and grouped holdout

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 is realistic settings.

Protocol	Majority Class	Convolutional NN
Uniform Holdout	7.8%	$37.58\% \pm 0.7\%$
Grouped Houdout	7.8%	$18.5\% \pm 1.3\%$

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.

Compare whole dataset with reduced dataset

While looking into the raw data clips, we noticed possible difficulty to distinguish very short clips. Thus, We removed clips shorter than 3 seconds from the dataset. Remaining 9723 clips (see Table 3. The grouped holdout protocol and the same model were used to fit the reduced dataset. As shown in Table 4, compared to the full dataset, the accuracy slightly decreased to 17.75% with a higher majority class baseline reference of 7.9%.

Table 2. Food types with break-down of number of videos and total number of clips taken from the videos. The full and reduced datasets can be compared.

Food Type	Videoclip Full	Videoclip Reduced
Aloe	12 - 547	10 - 158
Burger	12 - 596	11 - 182
Cabbage	12 - 500	9 - 445
Candied fruits	12 - 807	12 - 769
Carrots	12 - 661	12 - 622
Chips	12 - 720	12 - 701
Chocolate	13 - 291	13 - 279
Drinks	11 - 293	9 - 251
Fries	12 - 645	12 - 600
Grapes	12 - 580	12 - 552
Gummies	12 - 679	12 - 652
Ice-cream(coated)	14 - 728	14 - 711
Jelly	13 - 443	13 - 416
Noodles	13 - 412	13 - 363
Pickles	13 - 873	13 - 823
Pizza	12 - 610	12 - 604
Ribs	12 - 489	11 - 475
Salmon	14 - 502	14 - 481
Soup	12 - 279	11 - 160
Chicken wings	11 - 505	11 - 458

Table 3. Number of videos - clips for the full and reduced dataset(removed clips shorter than 3 seconds

Protocol	Majority Class	Convolutional NN
Full Dataset	7.8%	$18.5\% \pm 1.3\%$
Reduced Dataset	7.9%	$17.75\% \pm 1.2\%$

Table 4. Accuracy of the CNN using the grouped evaluation protocol on both full and reduced datasets. 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 3 shows the results of each pairwise classification task. The overall performance appears promising as the majority accuracy results are higher than 70% (visible as blue dots). The average accuracy and standard deviation per food type is shown in Figure 4. 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 sounding more alike. However, the unbalanced nature of the various problems might be a confounding factor. Figure 5 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] in which the clusters were mostly correlated to food textural differences.

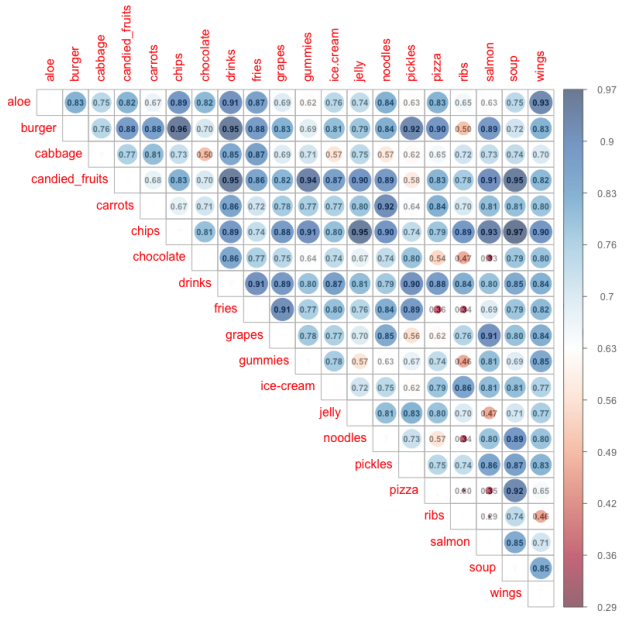

Figure 3. Pairwise classification result (accuracy). The numbers in each cell is the accuracy result of the pairwise classification task on the intersecting food type column and row. The scale shows the color coding of the accuracy values.

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. Since the experiment with

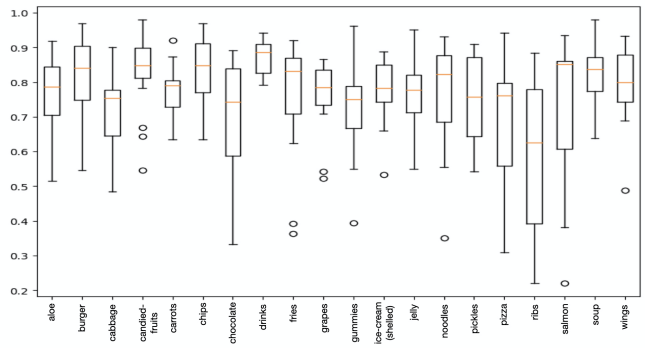

Figure 4. Boxplots of the average pairwise classification accuracy per food. If the food are classification with high accuracy from other food types, the average value is higher. Smaller standard deviation indicates being either outliers or very general among all food types.

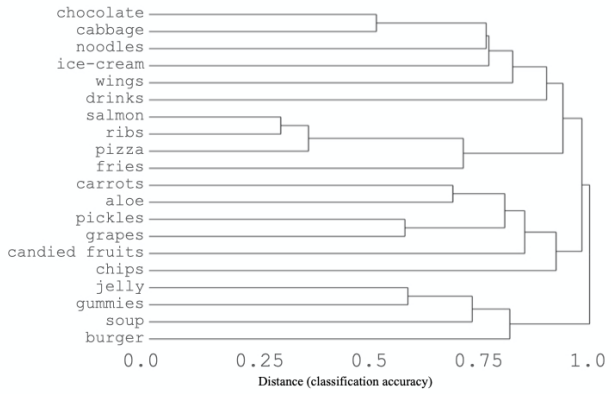

Figure 5. The dendrogram uses the classification accuracy as distance. Clustering appears less aligned with food textural differences compared to [1]

uniform holdout protocol outperformed the grouped evaluation protocol, the model might have learnt video differences as one of the major features. The reduced dataset did not improve the accuracy of classification. This might due to less learnable data resulted from removing short clips.

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 20-way classification task. 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 noises. The performance appears less accurate after balancing the results with the majority classes. The unbalanced file numbers of the dataset might have influenced as featured and learnt by the model for classification. From the boxplots we see drinks and candied fruites could be outliers among the group since it is generally easier to be classified aganist all other food types, and the standrd deviation is relatively smaller.

The dendrogram result (Figure 5) shows some relavance of textural differences among the classified food types. For example, crunchy vegitables like carrots, aloe, pickles and grapes are closely clustered as expected. However, some food types with similar texture appeared to be very different in this clustering result(e.g. drinks and soup). It indicated that there might be other audio clues featuring different food types rather than solely texture differences. As eating behaviour of different food can be much more than biting, food with similar texture might generate very different sounds while consuming (e.g. have drinks with a beaker of have soup with bowl and spoon). The previous works mainly focused on the biting sound which might have missed the most important features for certain food. Our result shows the biting sound classification is a limited protocol for food sound classification when the food types are not classified by texture types.

6. LIMITATION

Data related

- The clip separation methodology used in this paper was aimed to avoiding unwanted noise, but might have lost important feature clues in some clips compared to the others.
- The clip duration for different food types are unbalanced (e.g. aloe has a lot more shorter clips). After compressing to the same image size, the feature learnt by model could be less dependent to the sound features. Further normalizing the clips to similar length could be valuable to experiment on different length of clips.

Model related

- The model parameters (e.g. drop out rate) were inherited from previous researches on urban noises. However, the setting might not be exactly prefered for eating sound classification. The model performance might have room for improvement with hyperparameter tuners such as random searching.
- In this project, we used holdout method for spliting the training and validation data. The cross-validation method might provide different results though takes more time for training.
- For the pairwise classification, the unbalanced nature of the various problems might be a confounding factor for the accuracy differences. A balancing set of training dataset might help to improve the reasoning of clusters.

7. CONCLUSION

This research evaluated the performance of convolutional neural networks on food eating sound classification with online public-sourced training data, representing various real-life eating conditions, behaviours and recording qualities. As part of this study, eating sounds of 20 different food types was collected, processed, 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 trained with public sourced eating sound could only identify certain food from the 20 categories by 18.5% accuracy. With the uniform protocol, the model achieved 37.58% accuracy, indicating that the model might have learnt video patterns 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. The cluster of food shows separation of different textural composition for most of the food types. A few pairs of food with similar texture but different eating actions were distinctly separated. As an initial baseline, these results indicated the necessity of heuristic consideration while studing eating sound. The existing experimental protocols focusing on biting sounds might eliminate important sound cues present in real-world scenarios. With the aid of public data source, more features of eating-behaviour-related sounds could be introduced into the dataset to enrich evidences, thus, to improve the classification performance. Their inclusion might lead to better sound classification results for the purposes of classifying food on the basis of sound in an uncontrolled environment. Furthermore, the model's interpretation of eating sound could be experimented with generative methods like GANs. This project acts as a baseline and support for the possibel future researches on open-sourced data-driven eating sound classification.

8. ACKNOWLEDGEMENTS

I would like to thank my supervisors Dr. van Rijn and Mr. Gómez-Maureira for their great advices and support. While allowing full design freedom to me, they provided insightful suggestions which encouraged me towards continious improvement of my work, and inspired me with positive attitude on research. I would also like to thank the patient assistance from the helpdesk of The Academic Leiden Interdisciplinary Cluster Environment (ALICE). I would like to specially acknowledge the Media Technology chair group, my classmates, friends and family for connecting discussions and supporting eachothers during my thesis period under the situation of social distancing. Finally, I must express my gratitude to all the YouTubers for creating the public accessiable videos (snapshots displayed in Appendix A), contributed the bases of my thesis dataset.

9. REFERENCES

1. Oliver Amft. 2010. A wearable earpad sensor for chewing monitoring. In *Proceedings of IEEE Sensors Conference*. IEEE, 222–227.
2. François Chollet and others. 2015. Keras. `https://keras.io`. (2015).
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, and others. 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.

APPENDIX

A. SNAPSHOTS OF YOUYUBE VIDEOS SELECTED IN THIS PROJECT

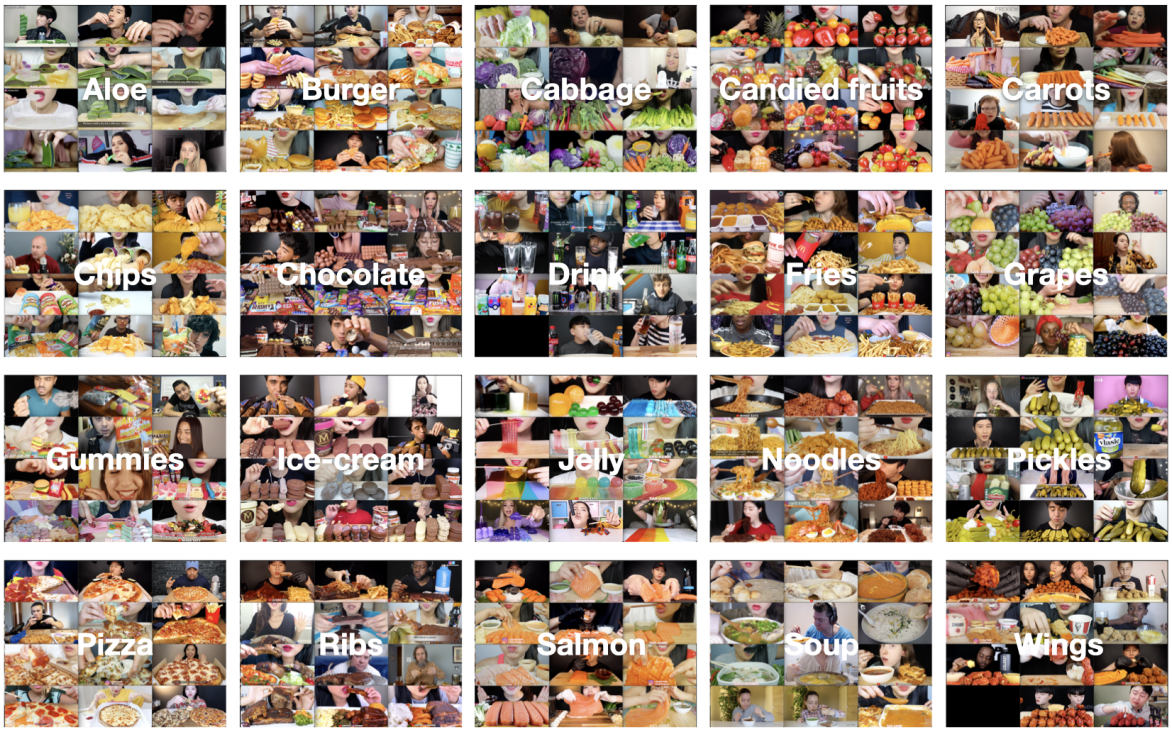

Figure 6. A glance of all the YouTube food eating videos used to create the dataset of this project grouped by food types.

B. PROJECT DEVELOPMENT

The objective topic for this thesis project has developed during the process of implementation. This section described the original proposal and insights for further researches.

B.1 Original Proposal

This project originally aimed to train a generative adverserial networks (GANs) model to generate eating sound using the collected dataset. The GANs model reflects Richard Feynman's quote: "what I cannot create I do not understand." It might be a good way to visualise the computer's perception and understanding of eating sound by giving it a chance to create them. When a random noise based generator compete with the judge who was trained by our dataset. We are curious to see how ambiguity and certainty valued in the generated sounds from this model. If the generation achieves promissing results, the model can be used for generating data for enforced learning on eating behaviour sounds, as well as providing possibility for automated virtual eating sound generation.

B.2 Objective Switch

The evaluation part of the GANs proposal is critical. In order to tell whether the generated sound is 'real' enough, a third party evaluator should be envolved. Applying human evaluation is one of the methods, but requiring unpredictable time and effort under the situation of this project. Therefore, we proposed to apply a trained classifier on both the raw dataset and the generated dataset. The difference of classification performance will be used to evaluate the performance of the GANs model. The classification of the raw dataset evolved many discussable results, which led us to a focus on the classification part for more detailed information of the raw dataset. Thus, the original proposal of GANs was splited out from this project.

B.3 Possible Further Researches With GANs

Throwing back to our inspiration, with the preliminary classification results from this project, further researches can proceed to test GANs on their reproducing ability of the eating sound they heard from human-food conversation. Based on our results, the GANs experiment is practical on some food pairs with significant differencs(higher classification accuracy). While fed with all the 20 types of eating sounds, the model is hypothesised to have a better chance for being ambigious, since the judge will have a difficult time to distinguish different food types and could tend to accept a more general sound.

C. TRIAL AND ERRORS

We did two trial experiments on part of the dataset to experiment with the model setting and get a sectional view of the dataset. This section reports the protocols and results of these trials:

C.1 Classify food types with data of 1 to 2 videos per food type

We trained the model with data from only 1 or 2 videos per food type using the protocol and parameter settings of the refered benchmark work (urban noise classificaiton). The train and validation data was randomly separated (7:3). The accuracy result for one trial is 0.84. This indicated that the model performed good classfication using samples from within 2 videos. This finding inspired us to investigate the influence of different video sources.

C.2 Classify videos within the dataset of one food type (aloe)

In this experiment we aimed to train the network to classify 12 different video sources for the clips from the aloe category. The train and validation data was randomly separated (7:3). Data from the training set and validation set may come from the same video. The test achieved an accuracy of 0.69. Compare to the guessing probability 15.4%, this result indicated a relatively significant influence of video source difference. Which may contribute to the performance of food type classification, as the network was allowed to learn patterns from different videos instead of different foods.

D. PARALLEL TASK SETTINGS

The Academic Leiden Interdisciplinary Cluster Environment (ALICE) was used to improve the efficiency of training. For the 20-way classificaiton task, the same 10 parallel jobs were submitted. For the pairwise classification task, folders containing corresponding pairs were arranged in advance. The 190 different pairs were splited into 2 jobs with array 0 99 and 100 189 to run in parallel.