Internship Report

Comparative Study of Classical Machine Learning Algorithm and RNN-LSTM for Estimation of Eating Activity from Chewing Sounds

Supervisors
Guillaume Lopez
Professor, Aoyama Gakuin University, Tokyo, Japan
Fabrice Muhlenbach
Associate Professor, University Jean-Monnet, Saint-Etienne, France

Submitted by- Shamprikta Mehreen

Abstract

Obesity is a medical problem that increases the risk of deadly diseases. Increasing the number of chews of each bite episode of a meal can help reducing obesity. However, it is really difficult for someone to keep track of his eating behaviors in detail without any help of an automatic mastication computing device. Even though such devices exist, they are not convenient for daily use. In the previous works, a cheap bone conduction microphone was introduced and it was used to collect sound data of different eating activities. Different machine learning methods have been proposed to classify different eating behaviors from the collected sound data correctly. In this research, a comparative study will be done for the classification of different eating behaviors. Study will be focused on comparing the classification of the eating activities such as, chewing, swallowing (food and drink) and talking using an RNN-LSTM model and a classical machine learning algorithm.

Contents

List of Figures

1 Introduction

Obesity can increase the risk of developing many potentially serious health conditions, including high blood pressure, diabetes and asthma. Most countries around the world regard obesity as a big problem, and the Japan Ministry of Health, Labor and Welfare is also trying to tackle this problem efficiently. In Japan, however, the number of obese people, that is with BMI 25 or more, has not decreased since ten years ago [1]. The relation between chewing and obesity is increasing number of mastication can help preventing obesity. Chewing well induces more saliva secretion and faster blood-sugar level increase. As a result, it works on the satiety center and hungry feeling is satisfied. That leads prevention of obesity [2]. As a concrete example, when attempting to improve mastication activity for young Chinese men with obesity, it was possible to reduce the intake of energy in all the subjects consistently [3]. Making conversation during meals is also related to health [4]. If the number of chewing and conversation during the meal can be detected, it is possible to give feedback to people in real-time, leading to prevent obesity. Therefore, the purpose of this research is to be able to classify the eating behaviors in a natural meal environment accurately in detail. Over the years, several progressive works have been done to classify the eating behaviors, including machine learning techniques. In this research, the goal is to compare an RNN-LSTM model with a classical machine learning algorithm for classification to estimate the eating activities from chewing sounds.

2 Related Works and State-of-the-Art

This section describes the state-of-the-art by presenting the related works and developed technologies and devices to quantify mastication activities. More than a decade ago, Amft et al. analyzed chewing sounds with a microphone placed inside the ear to enable getting high-quality chewing sounds [5]. They proposed to use mainly devices that measure myoelectric potential from the masseter muscle can count bites. However, wearing the apparatus in daily life is a significant burden for the user. The smart eyeglass was suggested by Zhang et al. for monitoring meals to sense chewing food, where they proposed a smart eye wear which contains a micro-controller and ElectroMyoGram electrodes, it was fabricated by a 3D printer [6]. Shuzo et al. analyzed eating habits with an IC recorder using a bone conduction microphone [7]. They calculated the power spectrum from audio data by using FFT (Fast Fourier Transform), then classified into four types of status - eating hard food, eating soft food, drinking water, and speaking - by features like the maximum frequency gained from the power spectrum. However, it is difficult to handle because of using a PC for analyzing and a bone conduction microphone whose sampling rate is 44.1kHz. Also, a bone conduction microphone connected to Bluetooth is required to handle it in everyday life. Therefore an algorithm that adjusts the sampling rate to 8kHz is necessary because many devices with Bluetooth connection have a sampling rate of 8kHz.

Bi et al. suggested a wearable device "Auracle", that can recognize meal behavior automatically [8]. The Auracle contains a a micro-controller, contact microphone, and an analog circuit connecting them.The contact microphone is placed behind the ear. A necklace-type device was suggested by Chun et al. for detecting meal activity [9]. It was created with a proximity sensor, a BLE (Bluetooth Low Energy) module, and a micro-controller. The movement is captured by the proximity sensor by distance to the jawbone and it can differentiate meal-related behaviors from non-meal-related behaviors. Mitsui et al. suggested a system that judges the number of chewing and the status of speaking in real-time by using a bone conduction microphone and gives real-time feedback to the user to improve his/her eating behavior [10]. However, they did not evaluate the performances in the natural meal environment yet, and the only used specific foods, such as onigiri (Japanese rice ball) and cabbage salad. Besides, the utterance method was a response to questions of the experimenter, which is not a natural type of conversation during a meal.

As summed-up above, despite many efforts by researchers over the last decade, an usable method for detailed tracking of dietary intake behavior in natural meal environment remains unsuccessful, and there is still room for improvement in judging detailed meal-related activities such as mastication amount per bite, speaking duration, and intake content. Kondo et al. collected dietary sound data by using a bone conduction microphone in a natural meal environment and classified into three eating behavior: chewing, swallowing and speaking [11]. They used SVM(Support Vector Machine) as a learning model fed with 75 features like Amplitude Difference Accumulation (ADA) and Short term energy (STE) for characterization of the signal. In the previous work, Jain et al classified different eating sound, adding noisy sounds collected in a natural meal environment [12]. In this research, along with the previous sound data used by Jain et al, some more sounds were added to classify different eating behaviors such as chewing, drinking, swallowing and talking. Classification was done using deep learning technique RNN-LSTM and a classical machine learning technique to compare their performance classifying different eating activities.

3 Collection of Daily Meal Sound

3.1 Data Collection

Dietary sound data was collected in a natural meal environment, with no restriction about meal content or surroundings. For example, some data were collected in a dining room and a standard household table with other family members, or at the university cafeteria with friends, such we can assume that represents different noisy conditions. The meal content was also totally free, and subjects ate whatever they wanted as usual in daily life, such various food types were mixed unpredictably. In our previous work, we labeled each sound data sequence corresponding to a single chew, swallow, speaking events. In this work, new data were collected in that situation to existing data, and that increased the number of samples for each label.

To collect dietary sound data, a commercial bone conduction microphone (Motorola Finiti HZ800 Bluetooth Headset, Motorola co. Ltd.) was used, attached to one ear of the subject, that can operate Bluetooth communication with a smartphone (Google Pixel 3, Google co. Ltd.) and collected dietary voice data using a dedicated Android OS application. The sound signal sampling from the microphone was 8KHz. Dietary sound data was collected from five men and four women aged between 11 and 23 years old, for a total of 13 meals, five different meals for subject 1, and one meal for each of the other eight subjects. After collection, data were transferred to a computer for analysis. Besides, since data were collected in a totally free environment, it was necessary to perform labelling afterwards. To label sound segments corresponding to chewing, swallowing, and speaking, after collecting the data, a video was taken together with sound data to assist the labeling work. The video shooting was performed so that the mouth and throat of the subject were reflected. Figure 1 shows a picture of the data collection conditions (for privacy, it is a photograph that reproduces the actual environment).

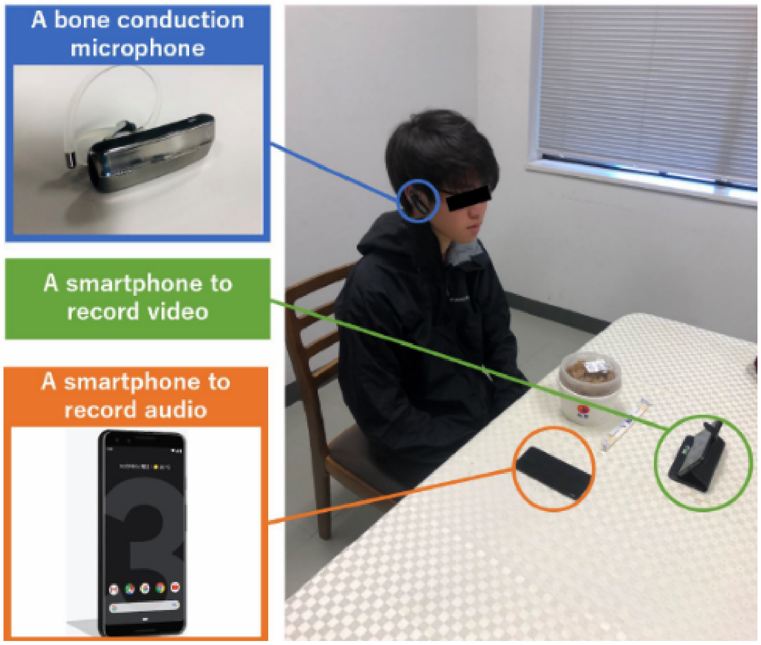
A bone conduction microphone
A smartphone to record video
A smartphone to record audio

Figure 1: Picture reproducing the data collection conditions

3.2 Arrangement of Data Labelling

It was necessary to arrange the audio data for labelling that were collected in a natural meal environment as described in the previous section. This process aimed to make ground truth dataset used to build-up a machine-learning model that can classify four eating behaviors: chewing, drinking, swallowing and talking. Differentiation of food and drink is a crucial issue to separate different bites and enable further detailed eating habit quantification. The audio sections corresponding to each eating behavior were labeled to make the ground truth dataset.

While labeling audio data collected by the bone conduction microphone, video taken simultaneously with the recording of the audio data was also used because it is difficult to label meal-time chewing, drinking, swallowing and talking using only audio data. Figure 2 shows a screenshot of the synchronization process of collected audio data and video by using Adobe Premiere Pro CC (produced by Adobe Systems Incorporated). The sound of the taken video was replaced with the audio data collected by the bone conduction microphone by synchronizing audio data and the video.

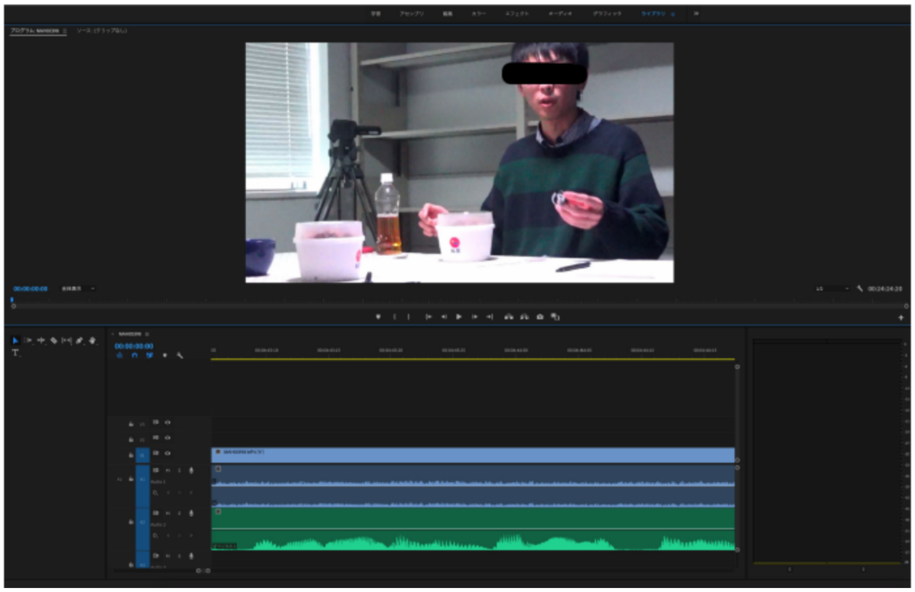

Figure 2: Screenshot of the application software used to synchronize audio and video

3.3 Data Labelling

Audio data were labeled using Praat [13], which is an audio analysis software that can associate labels to audio data sections. The labels were set according to the targeted four eating behaviors: “chewing” (C), “food swallowing” (S), “drink swallowing” (Drink), “talking” (T), and “other” (O). Labeling was performed by referencing the raw data and intensity of the audio signal and the synchronized video. Figure 3 shows the audio data labeling using Praat.

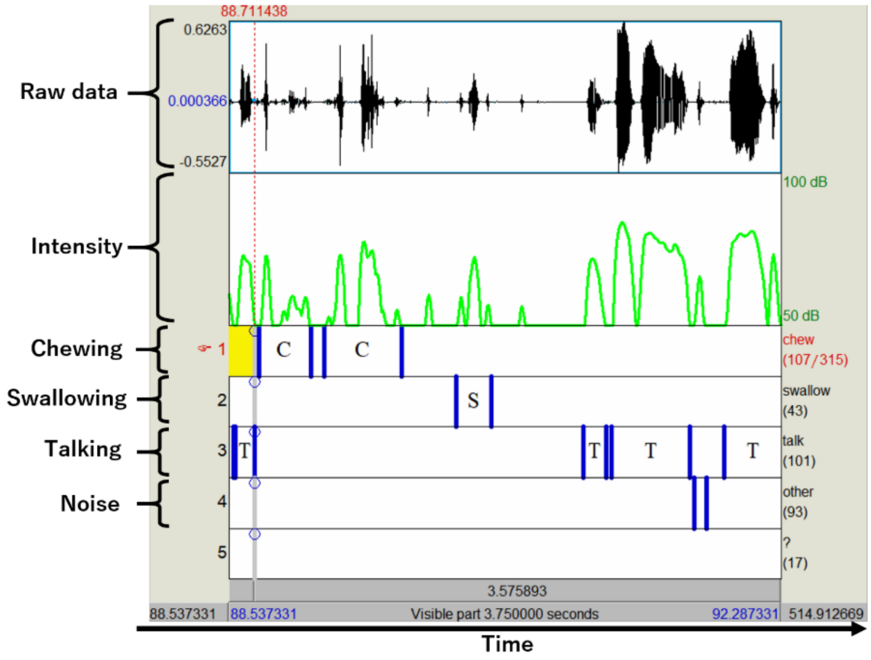

Figure 3: Audio data labeling with Praat

Audio sections were extracted into segments according to their label. Figure 4 shows examples of raw data of the extracted audio segments of a chew(top left), a drink(top right), a swallow(bottom left) and a talk(bottom right) segment.

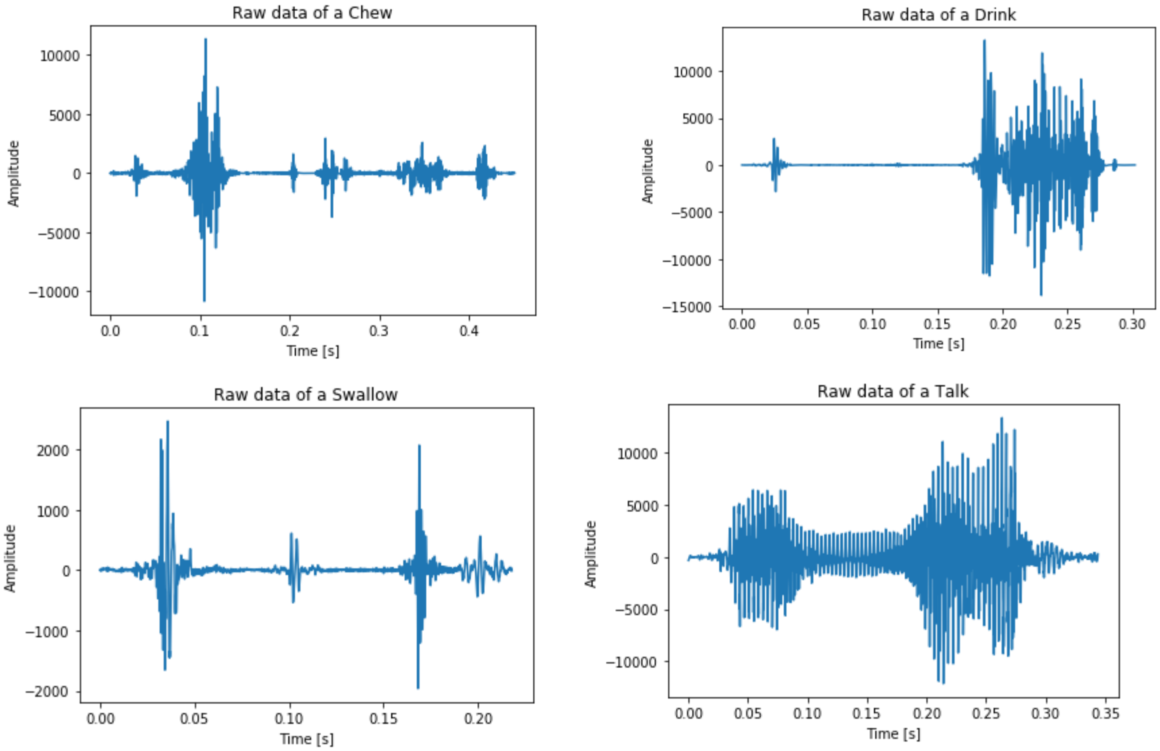

Figure 4: Raw data of one chew segment(top left), one drink segment(top right), one swallow segment(bottom left), one talk segment(bottom right)

Following the above described procedures a dataset was prepared. The resulting dataset is described in Table 1. The dataset represents 144 minutes of 3232 effective eating sound segments from unconstrained meals of various types of food, resulting in 2245 chewing samples, 151 drinking samples, 143 swallowing samples and 693 talking samples.

S	Nb of Chew	Nb of Drink	Nb of Swallow	Nb of Talk	Total meal time [min]
1	152	5	12	88	18:12
2	183	0	11	55	7:25
3	160	8	10	28	12:08
4	208	0	8	65	8:21
5	486	14	9	55	5:43
6	188	0	17	50	3:17
7	144	4	9	16	8:58
8	111	6	9	103	5:33
9	91	0	4	13	3:54
10	203	0	14	0	4:28
11	157	0	2	9	8:42
12	108	11	26	71	8:39
13	162	5	17	57	10:07
14	216	12	4	30	32:48
15	0	4	0	4	0:35
16	0	6	0	7	0:41
17	0	5	0	14	0:38
18	0	1	0	1	0:29
19	0	18	0	10	1:09
20	0	10	0	2	0:32
21	0	7	0	12	1:00
22	0	6	0	0	0:14
23	0	13	0	3	0:21
24	0	14	0	0	0:25
Total	2245	151	143	693	144:08

Table 1: Detail of the amount of labels and data sections extracted for each meal related activity type

4 Data Preprocessing

4.1 Feature Extraction

In this research, Features extraction has been performed from the dataset labeled according to the previous section before operating machine learning models for meal-related activities classification. A total of 26 features were extracted. Table 2 shows the name and number of extracted features.

Features	Number of Features
Chroma Vector	1
Spectral Centroid	1
Spectral Bandwidth	1
Spectral Rolloff	1
Root Mean Square Energy (RMSE)	1
Zero Crossing Rate	1
Mel Frequency Cepstral Coefficients (MFCC)	20

Table 2: Extracted 26 features Features

4.1.1 Chroma Vector

A chroma vector is a typically a 12-element feature vector indicating how much energy of each pitch class C, C# , D, D#, E, F, F#, G, G#, A, A#, B is present in the signal. One main property of chroma features is that they capture harmonic and melodic characteristics of music, while being robust to changes in timbre and instrumentation. It is also used for audio- matching making[14].

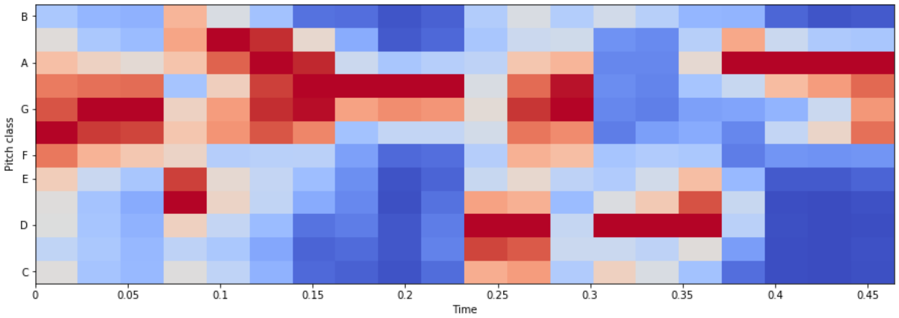

Figure 5: Chroma Feature extracted for a single chew segment

4.1.2 Spectral Centroid

The spectral centroid indicates where the ”centre of mass” for a sound is located and is calculated as the weighted mean of the frequencies present in the sound. If the frequencies in music are same throughout then spectral centroid would be around a centre and if there are high frequencies at the end of sound then the centroid would be towards its end[15]. This is like a weighted mean:

$$f_c = \frac{\sum_k S(k) f(k)}{\sum_k S(k)} \tag{1}$$

Where $S(k)$ is the spectral magnitude at frequency bin k, $f(k)$ is the frequency at bin k.

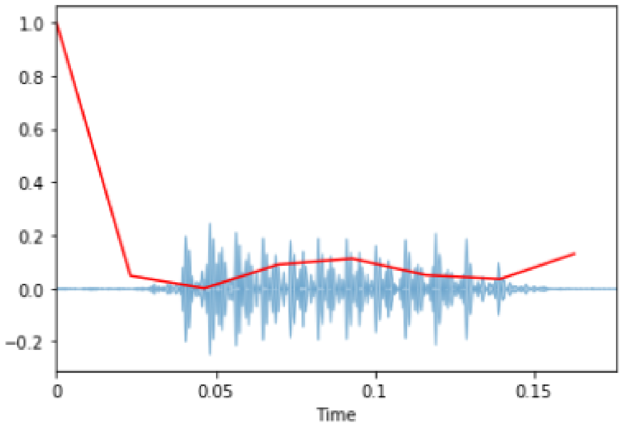

Figure 6: Spectral Centroid extracted for a single talk segment

As we can see from figure 6, there is a spurious rise in spectral centroid at the beginning of the segment. That happened because the silence at the beginning has such small amplitude that high-frequency components have a chance to dominate.

4.1.3 Spectral Bandwidth

It computes the order- p spectral bandwidth:

$$(\sum_{k} S(k)(f(k) - f_c)^P)^{\frac{1}{P}} \tag{2}$$

Where $S(k)$ is the spectral magnitude at frequency bin k, $f(k)$ is the frequency at bin k, and f_c is the spectral centroid. When $p = 2$,this is like a weighted standard deviation.

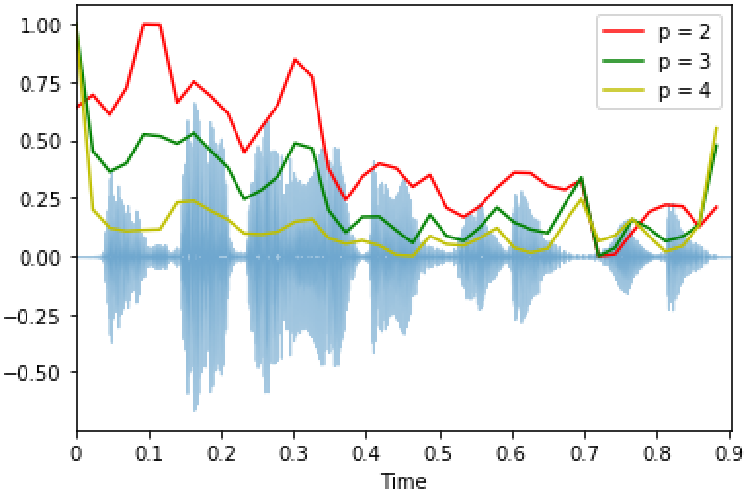

Figure 7: Spectral bandwidth extracted for a single chew segment

4.1.4 Spectral Roll-off

Spectral Roll-off is a measure of the shape of the signal. It represents the frequency at which high frequencies decline to 0. To obtain it, we have to calculate the fraction of bins in the power spectrum where 85% of its power is at lower frequencies [16]. It also gives results for each frame.

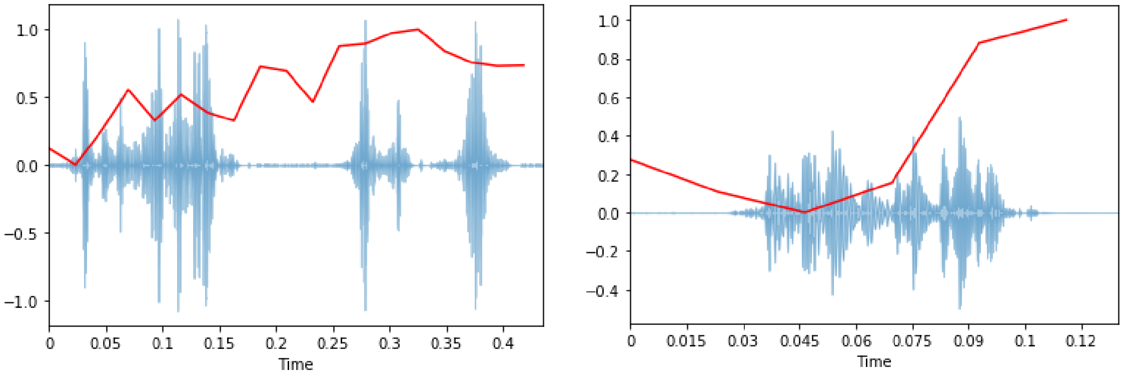

Figure 8: Spectral Roll-off extracted for a single drink segment(left) and a single swallow segment(right)

4.1.5 Root Mean Square Energy (RMSE)

The energy of a signal corresponds to the total magnitude of the signal. For audio signals, that roughly corresponds to how loud the signal is [17, 18]. The root mean square energy (RMSE) of a signal segment s containing N samples is defined as the square root of the average of the sum of all samples n:

$$\sqrt{\frac{1}{N}\sum_{n}|x(n)|^2} \tag{3}$$

4.1.6 Zero Crossing Rate

The zero crossing rate indicates the number of times that a signal crosses the horizontal axis [19]. It is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to zero to negative or from negative to zero to positive[20]. The zero crossing rate indicates the number of times that a signal crosses the horizontal axis.

4.1.7 Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients(MFCC) are the he most used feature for speech recognition. They capture timbral/textural aspects of sound. The great advantage of MFCCs is they approximate human auditory system, i.e, they try to model the way human beings perceive frequency. The result of extracting MFCCs is a bunch of coefficients(usually between 13 to 40). They are calculated at each frame.

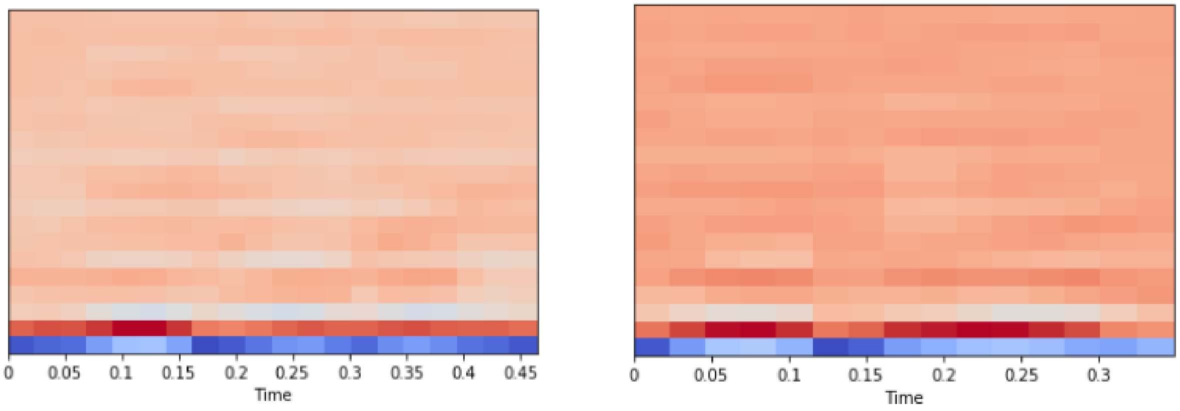

Figure 9: 20 MFCCs extracted for a single chew segment and a single swallow segment. The vertical axis represents the MFCC coefficient order and the color represents the value of each coefficient (red: small to blue; big).

After extracting the features, each feature value was scaled such as the average is zero and the variance is one, by using StandardScaler. Also, the four categorical labels-chew, drink, swallow, talk were converted into model-understandable numerical value using Label Encoder class.

4.2 Train and Test Set

The whole data set was randomly splitted in a stratifying fashion, using our labels into two sets, 80% as training set and 20% as test set. Table shows the number of samples in each label after splitting the dataset.

Label Name	Total	Train	Test
chew	2245	1796	449
drink	151	121	30
swallow	143	114	29
talk	693	554	139
Total	3232	2585	647

Table 3: Number of samples for each activity label after splitting the dataset into training and test set

4.3 Balancing the imbalanced dataset

As we can see from the table above, our dataset is highly imbalanced. The number of drink and swallow is really low compared to the number of chew and talk. Which is a big problem since the goal is to classify the labels correctly. To deal with this issue, an oversampling method called SMOTE(Synthetic Minority Oversampling Technique) has been applied to balance the dataset. Other oversampling method such as ADASYN was also tried, and it gives the similar result as SMOTE. Oversampling can be done before splitting it into training data and test data which makes the the model perform really good. But in this case, the test data will not be real world data, which we do not want, and it might lead the machine learning model to overfit. Keeping this problem in mind, the oversampling method has been applied only on the training data. Table shows the number of samples in each label after applying SMOTE on training set.

Label Name	Train	Test
chew	1796	449
drink	1796	30
swallow	1796	29
talk	1796	139
Total	7184	647

Table 4: Number of samples for each activity label after using SMOTE

5 Meal-Time Activities Classification

5.1 RNN-LSTM for classification

LSTM Networks

Long Short Term Memory(LSTM) networks are a special kind of Recurrent Neural Network, capable of learning long-term dependencies. The LSTM has the capability to remove or add information to the cell state(C), carefully controlled by gates. Gates are composed of sigmoid neural net layer and a multiplication operation,they are a way to optionally let information through. The sigmoid layer gives the ooutput between 0 and 1, while 0 indicates to not to let the information pass, and 1 indicates the opposite. An LSTM has three of these gates to protect and regulate the cell state. The first sigmoid layer controls which information to forget. The second one controls which information to remember/store in the cell state, which then combines with a tanh layer to create an update to the state. The third sigmoid layer decides which parts of the cell state is going to be the output. Finally, the cell state is passed through a tanh(to push the values between -1 and 1), which gets multiplied by the output of the sigmoid gate. So, it outputs only the part it has decided to.

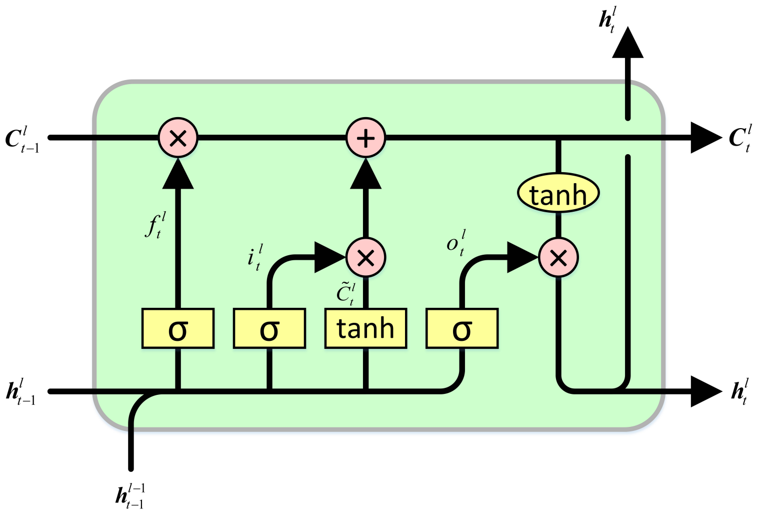

Figure 10: LSTM cell

Here in our Meal-Time activities Classification, we have used the LSTM to classify different eating behaviors, such as Chewing, Drinking, Swallowing and Talking.

5.1.1 Choosing the right Hyperparameters

Reshaping data for feeding into LSTM

The input shape of every LSTM layer must be three-dimensional. The three dimensions of this input are: Samples, Time steps and Features[21].

- **Samples**: One sequence is one sample.
- **Time steps**: One time step is one point of observation in the sample.
- **Features**: One feature is one observation at a time step.

We had 3232 samples with 1 time-step and 26 features. We reshaped the training data in such a way that our LSTM model gets a 3D input shape with 1 time-step and 26 features. After our LSTM layers did all the work to transform the input to make predictions towards the desired output possible, we have to reduce (or, in rare cases extend) the shape, to match our desired output[22]. In our case, we have four output labels (chewing, drinking, swallowing, talking) and therefore we need four-output units.

A big part of deep learning is Hyperparameter optimization. The reason behind this is neural networks are difficult to configure and there are a lot of parameters that need to be set. Also, some models can be very slow to train. For our LSTM model, Grid Search model hyperparameter optimization technique was used. Several hyperparameters were tuned, such as number of Epochs, Batch Size, Learning Rate and Dropout Regularization. The best epoch was 50, batch size 45, learning rate 0.0001 and dropout rate 0.4. 3-fold cross validaition was used for evaluation.

For activation functions, 'softmax' was used in the output layer because it allows the model to interpret the outputs as probabilities for each class. Sparse Multiclass Cross-Entropy Loss was used as loss function. Optimizer 'Adam' was used and finally 'F1 Score' was used as metrics to judge our model.

Overfitting
As deep neural networks are very prone to overfit, so in order to prevent overfitting, Dropout layers were added in the model. Also, overfitting was diagnosed in the LSTM model by plotting the training loss and testing loss over epochs. Overfitting occurs when a model has learned the training dataset too well, including the random fluctuations or statistical noise from the training dataset. An example is shown in figure 11. As we can see from the figure, the training loss continues to decrease over epochs and the test loss decreases to a point and then begins to increase again, which demonstrate a case of overfitting. In our case, we tried to deal with this problem by putting the best dropout rate(which we got from Grid Search hyperparameter optimization) in the dropout layer.

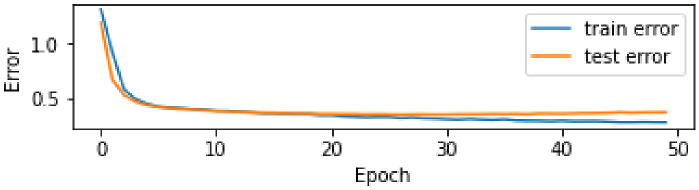

Figure 11: Diagnosing Overfitting

5.2 Classical Machine Learning Algorithms for Classification

To compare our LSTM model with a classical machine learning algorithm, Support Vector Machine(SVM) was used to classify the four different eating behaviors- chewing, drinking, swallowing and talking. Grid search method was used for optimizing SVM parameters "C" and "Gamma". These two parameters control the complexity of the model and increasing them results in more complex models. In addition, the setting of these two parameters is strongly correlated. Therefore, "C" and "gamma" must be adjusted at the same time. Six different values of "C" and nine different values of "Gamma" were tested using a five-fold cross validation. The best optimized model was obtained was rbf kernel, with "C" value 10 and "Gamma" value 0.1. Using the optimized parameters, the performance of the model was evaluated with the test dataset, which was 20% of the whole dataset. An F1-Score of 0.86 was obtained.

6 Results

6.1 Confusion Matrix comparison between RNN-LSTM and SVM

After training our LSTM model with 80% training data, model was evaluated on the test data and the F1-score of 0.86 was obtained. Comparatively, after training our SVM classifier with 80% training data and best 'C', 'Gamma' and RBF kernel, model was evaluated on the test data and the F1-Score was 0.86. Figure 12 shows the confusion matrices on test data after classifying four different eating behaviors using LSTM and SVM.

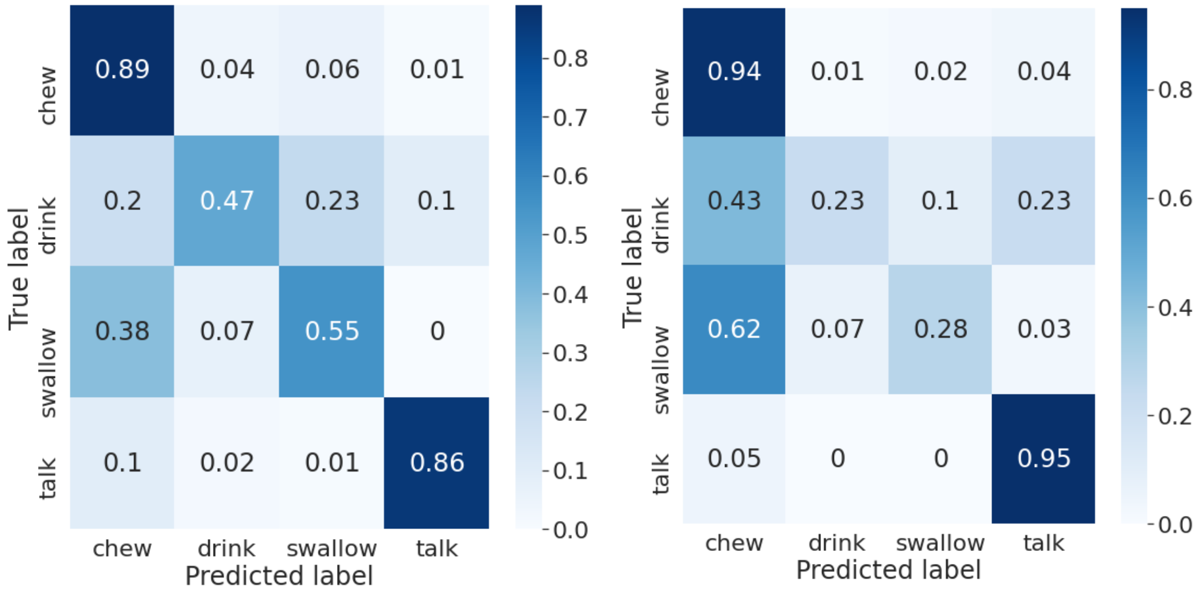

Figure 12: Confusion matrix between predicted label and true label on test data for LSTM (left) and SVM rbf model(right)

	precision	recall	f1-score	support
chew	0.93	0.89	0.91	449
drink	0.39	0.47	0.42	30
swallow	0.30	0.55	0.39	29
talk	0.93	0.86	0.90	139
accuracy			0.85	647
marco avg	0.64	0.69	0.65	647
weighted avg	0.88	0.85	0.86	647

	precision	recall	f1-score	support
chew	0.92	0.94	0.93	449
drink	0.47	0.23	0.31	30
swallow	0.44	0.28	0.34	29
talk	0.85	0.95	0.89	139
accuracy			0.88	647
marco avg	0.67	0.60	0.88	647
weighted avg	0.86	0.88	0.86	647

Table 5: Score from LSTM(left) and SVM rbf(right) model

From our result, it is certain that both the models can classify sound segments for chew and talk, but they cannot classify sound segments for drink or swallow properly. The reason behind this is the dataset is highly imbalanced as it has really less number of drink and swallow sound samples. Despite of having the same F1-score, the LSTM model seems to classify the samples for drink and swallow better than SVM classifier. From table 5, we can see the performance differences between LSTM model and SVM. We can see that recall is better with SVM model, but precision better with LSTM model.

6.2 Result after using SMOTE on both Training and Test data

Even though SMOTE oversampling method was applied, it was only on training data. So, to compare,we used SMOTE on both the training and testing data so that we do not face this problem. An F1-score of 0.97 was obtained by LSTM and 0.98 was obtained by SVM classifier. Figure 13 shows the confusion matrices of both LSTM and SVM, after applying SMOTE on both training and test data, and it shows a typical example of overfitting.

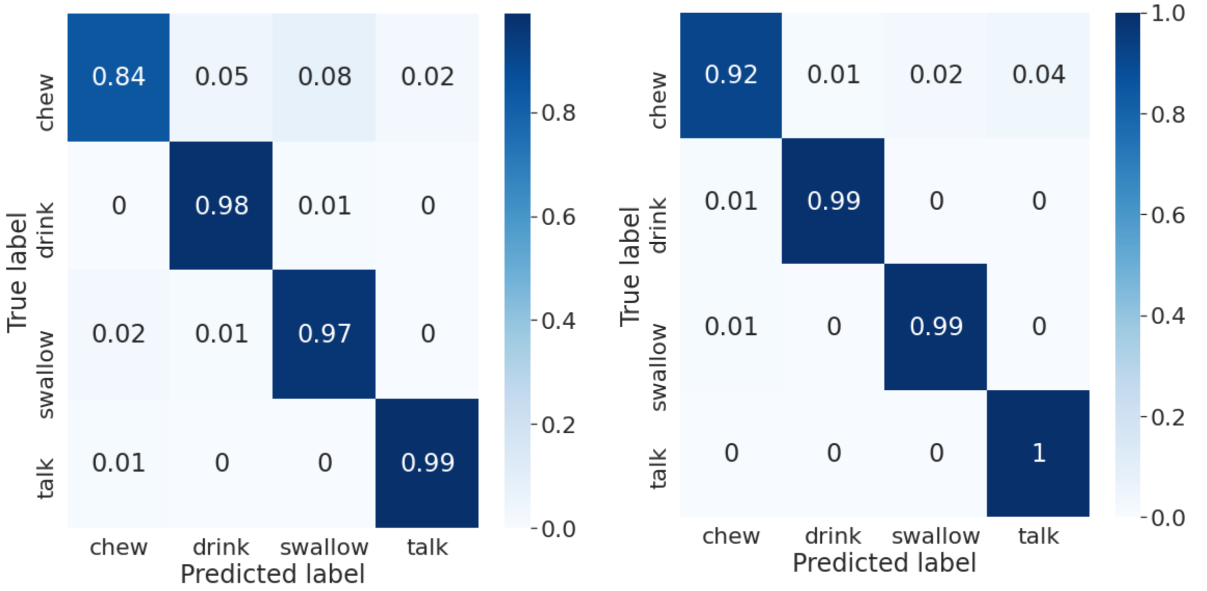

Figure 13: (After applying SMOTE on both training and test data) Confusion matrix between predicted label and true label on test data for LSTM(left) and SVM rbf model(right)

7 Challenges

It was shown in the previous section that both our RNN-LSTM and classical machine learning (SVM) model performed well while classifying the sound segments for chew and talk, which were more than 14 times bigger than the rest of the two eating behaviours (drink and swallow), so the dataset was imbalanced. To balance the imbalanced dataset, oversampling method SMOTE was used. Applying SMOTE on both training and test data gives better accuracy on test data. But the problem here is the test data is not real world data and we want to test our Model on real world data to find the true performance of the Model. Also, this leads to overfitting. So, oversampling method SMOTE was used only on the training data for both LSTM and SVM model.

8 Conclusions and Future Work

In this study, we compared LSTM model and classical machine learning model, Support Vector Machine(SVM) for classification using sound data for different eating behaviors. 26 features were extracted and classification of four eating behaviors(chewing, drinking, swallowing, talking) was done using LSTM and SVM(using Radial basis function kernel) separately. For both classifiers, hyperparameters were optimized using Grid Search method. Both the models performed well when the dataset was balanced and were able to classify the different eating behaviors, though they could not perform well enough classifying the sound segments for drinking and swallowing, because of the less number of samples for these two classes. They can achieve better performance classifying the dietary sounds if there are more samples for drinking ans swallowing sounds.

The future work should be to increase the samples for drinking and swallowing to make a balanced dataset. Feature engineering should be done and different oversampling methods can be used to increase the model performance. Automatic segmentation can be tried by LSTM model. In addition, estimation of the amount of texture (hardness, softness, crispiness, etc) of different food can be done.

References

[1] Japan ministry of health labor and welfare. the national health and nutrition survey in japan, heisei 29. `https://www.mhlw.go.jp/content/10904750/000351576.pdf`.

[2] Kanazawa medical association. about chewing. `http://www.kma.jp/ishikai/ishikai_0062.html`.

[3] Jie Li, Na Zhang, Lizhen Hu, Ze Li, Rui Li, Cong Li, and Shuran Wang. Improvement in chewing activity reduces energy intake in one meal and modulates

plasma gut hormone concentrations in obese and lean young chinese men–. *The American journal of clinical nutrition*, 94(3):709–716, 2011.

[4] Noriko Kishida and Yoshie Kamimura. Relationship of conversation during meal and health and dietary life of school children. *The Japanese Journal of Nutrition and Dietetics*, 51(1):23–30, 1993.

[5] Oliver Amft, Mathias Stäger, Paul Lukowicz, and Gerhard Tröster. Analysis of chewing sounds for dietary monitoring. In *International Conference on Ubiquitous Computing*, pages 56–72. Springer, 2005.

[6] Rui Zhang, Severin Bernhart, and Oliver Amft. Diet eyeglasses: Recognising food chewing using emg and smart eyeglasses. In *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 7–12. IEEE, 2016.

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

[8] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, et al. Auracle: Detecting eating episodes with an ear-mounted sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–27, 2018.

[9] Keum San Chun, Sarnab Bhattacharya, and Edison Thomaz. Detecting eating episodes by tracking jawbone movements with a non-contact wearable sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–21, 2018.

[10] Hideto Mitsui, Joe Ohara, Anna Yokokubo, and Guillaume Lopez. Method to improve real-time chewing and speaking detection accuracy from bone-conduction sound. *Method to improve real-time chewing and speaking detection accuracy from bone-conduction sound*, 2(1):1–21, 2018.

[11] Takumi Kondo, Hidekazu Shiro, Anna Yokokubo, and Guillaume Lopez. Optimized classification model for efficient recognition of meal-related activities in daily life meal environment. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 146–151. IEEE, 2019.

[12] Archit Jain, Takumi Kondo, Haruka Kamachi, Anna Yokokubo, and Guillaume Lopez. Detailed classification of meal-related activities from eating sound collected in free living conditions. In *HEALTHINF*, pages 284–291, 2020.

[13] Praat: doing phonetics by computer. `http://www.fon.hum.uva.nl/praat/`.

[14] Chroma feature. `https://en.wikipedia.org/wiki/Chroma_feature`.

[15] Music feature extraction in python. `https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d`.

[16] Notes on music information retrieval. `https://musicinformationretrieval.com/spectral_features.html`.

[17] Energy (signal processing). `https://en.wikipedia.org/wiki/Energy_(signal_processing%29`.

[18] Notes on music information retrieval. `https://musicinformationretrieval.com/energy.html`.

[19] Notes on music information retrieval. `https://musicinformationretrieval.com/zcr.html`.

[20] Zero-crossing rate. `https://en.wikipedia.org/wiki/Zero-crossing_rate#cite_note-1`.

[21] How to reshape input data for long short-term memory networks in keras. `https://machinelearningmastery.com/reshape-input-data-long-short-term-memory-networks-keras/`.

[22] Choosing the right hyperparameters for a simple lstm using keras. `https://towardsdatascience.com/choosing-the-right-hyperparameters-for-a-simple-lstm-using-keras-f8e9ed76f046`.