Classification Method of Eating Behavior by Dietary Sound Collected in Natural Meal Environment

Haruka Kamachi, Takumi Kondo, Anna Yokokubo, and Guillaume Lopez

Abstract Having a conversation during a meal, eating slowly, and thorough chewing, are some of the passive weight-loss strategies. Furthermore, detecting eating behaviors such as the number of chewing and the duration of the conversation leads to positive dietary behavior. This paper proposes a method that can accurately quantify eating behavior in a natural meal environment. We used a bone conduction microphone and recorded the dietary sounds of 16 subjects. We manually labeled five eating behaviors, namely chewing, swallowing food, swallowing drink, speaking, and other sounds like noise. We then extracted 75 features from the collected dataset and applied appropriate machine learning algorithms and categorized the eating behaviors. The resulting models of discriminating between chewing and speaking was possible with high F1 score. However, they achieved a lower accuracy in classifying swallowing and other sounds, especially swallowing food and swallowing drink. Furthermore, the machine learning models confused swallowing food and swallowing drink as chewing. Therefore, it is necessary to find features that express better the differences between these three behaviors.

1 Introduction

Obesity causes lifestyle diseases such as diabetes and high blood pressure. The Japan Ministry of Health, Labor and Welfare is trying to tackle this problem. However, the number of obese people has not decreased in the past ten years [1].

Eating too fast may lead to overeating because the brain does not have enough time to estimate one's fullness level and increases one's BMI [3]. On the contrary, slow eating promotes thorough chewing and induces the salivary glands to release

Haruka Kamachi
Aoyama Gakuin University, Sagamihara Japan, e-mail: hkamachi@wil-aoyama.jp

Takumi Kondo
Aoyama Gakuin University, Sagamihara Japan, e-mail: tkondou@wil-aoyama.jp

Anna Yokokubo
Aoyama Gakuin University, Sagamihara Japan, e-mail: yokokubo@it.aoyama.ac.jp

Guillaume Lopez
Aoyama Gakuin University, Sagamihara Japan, e-mail: guillaume@it.aoyama.ac.jp

more saliva. It also hasten the increase in blood-sugar level. As a result, slow eating reduces overeating and prevents obesity [2].

These days, solitary eating is a serious problem because of decreasing opportunities to eat together with one's family. Kishida et al. [4] have reported that making conversation during the meal is related to health. They showed that when there is a conversation with family during the meal, the lifestyle was regular and the proportion of healthy people was higher. These people tend to eat many vegetables and have fewer likes and dislikes in their dietary habits.

Recently, there exist many commercial wearable devices that enable monitoring people's activity level by measuring daily calorie consumption. However, none of these devices can detect automatically dietary behavior in a natural meal environment. Such a device would allow giving suggestions on the number of chewing, the pace of eating, and the length of conversations by making it possible to detect and with high accuracy chewing and speaking. It would lead to improving the consciousness of positive dietary behavior, such as increasing the number of chewing and developing conversations. A previous study showed that in a laboratory environment, real-time feedback about the number of chewing during the meal promotes the increase in the number of chewing[5].

This paper proposes a model to classify, in a natural meal environment, chewing, speaking, swallowing food, and swallowing drink.

2 Related Work about Automatic Dietary Activity Analysis

This chapter illustrates the state-of-the-art by presenting existing works and technologies related to automatic dietary activity detection.

2.1 Method for recognition and analysis of dietary activity from jaw motion

Zhang et al. [6] proposed a smart eyeglass for monitoring food chewing. The precision and recall of chewing activity detection from EMG signal analysis both achieved 80%. Also, they were able to classify five kinds of food with an accuracy in the range of 63% to 84%. Chun et al. [7] suggested a necklace-type devices to detect meal activity. It uses a proximity sensor to capture movement by distance to the jawbone and distinguishes between meal-related and non-meal-related behaviors. While this device showed excellent results for meal activity detection with a precision and recall of respectively 91.2% and 92.6% in a experiment environment, it resulted in more than 10% performance decrease in a free-living environment.

Wang et al. [8] suggested headband device for eating detection and chews counting. It uses triaxial accelerometer on the temporalis to obtain the bulge of the mastication

muscles. The average accuracy and F score for eating activity detection are 94.2% and 87.2%. Also, the average error rate of chews counting is 12.2%.

Bedri et al. [9] have developed a one-size-fits-all pair of eyeglasses that tracks when a user eats or drinks using an infrared proximity sensor, five gyroscope, and an accelerometer and aids identifying food type using a camera. They could recognize eating episodes with 94.1% accuracy and estimate the duration of the eating episode with 96.3%.

Although they could detect accurately eating activity, original devices and expert knowledge are necessary and detailed eating behavior was not analyzed.

2.2 Method for recognition and analysis of dietary activity from eating sound

Amft et al. [10] analyzed chewing sounds with a microphone placed inside the ear to enable getting high-quality chewing sounds. Meal activity detection accuracy was 99%, and the accuracy in classifying four kinds of food exceed 80%. Bi et al. [11] suggested a wearable device that can recognize meal behavior automatically by using a contact microphone. The contact microphone is placed behind the ear. This device can detect the meal activity in a free-living environment and could achieve high performances.

From these studies, a microphone is good for high accuracy classification of detailed eating behavior. Shuzo et al.[12] and Zhang et al.[13] analyzed eating activity with an IC recorder using a bone conduction microphone. A contact microphone such as a bone conduction microphone is less affected by ambient noise since it captures internal vibrations from the body surface directly. They classified into some types of food and texture, and the classification accuracy was high. However, experiment was in experimental environment and this device is not easy to use in a natural meal environment. Also, Mitsui et al. [5] 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. They could count in real-time chewing behavior with 91% accuracy and utterance length with 96% accuracy. However, they did not evaluate the performances in the natural meal environment yet, and the only used specific foods. Besides, the utterance method was a response to questions of the experimenter, which is not a natural type of conversation during a meal.

Mirtchouk et al. raised the significant gap between lab and free-living environments conditions [14]. Works highly accurate according to lab conditions, perform when tested in realistic environments. They presented a multi-modal study on eating recognition by combining lab-environment and free-living environment data collection. However, in "free-living" conditions, participants had to wear a pair of smart-glasses, an earbud, and a smartwatch on each wrist. Besides, participants were given a portable food scale for recording food weights at the start and end of each meal. Though there were no constraints in food content and time, the participants had

to follow some protocols (e.g., knocking the table to synchronize signals), and all these apparatus make the environment far different from daily. Moreover, the purpose of their analysis was to detect eating episodes combining multiple modalities. On the other hand, our study aims to detect specific eating activities (chewing, swallowing food and drink, etc.) from a single modality.

2.3 Classification Method of detailed eating behavior

Kondo et al. [15] 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. The precision, recall and F1 score exceeded 90%. Besides, in follow-up work [16], they extended the model to the classification of four activities - chewing, swallowing, speaking, and other like noise - and achieved more than 90% accuracy on average using only the best 7 selected features and best 14 features. However, oversampling of the unbalanced dataset was performed before dividing into training data and test data. In this case, data that not exist are used as test data, which may lead to overfitting the machine learning model.

2.4 Summary

As summed-up above, using a bone conduction microphone is suitable for detailed eating behavior detection. However, previous studies were not conducted in a natural meal environment and the machine learning models may have been overfitted because of incorrectly resampling timing. Oversampling of the unbalanced dataset has been performed before dividing it into training data and test data. In this case, data that does not exist are used as test data, which may lead to overfitting the machine learning model. Also, swallowing food was not distinguished from swallowing drink. Therefore, in this paper, we use a bone conduction microphone for classification of specific dietary intake behavior (chewing, speaking, swallowing food, swallowing drink) in a natural meal environment, and oversampling is performed after dividing it into training data and test data.

3 Classification Method

The entire experimental protocols got the approval of the ethics committee of Aoyama Gakuin University. The author of this research and the persons who collaborated to collect data have all completed the e-learning course of research ethics provided by the Japan Society for the Promotion of Science.

3.1 Collection of Daily Meal Sound

Dietary sound data was collected in a natural meal environment to classify chewing, two types of swallowing (food and drink), speaking, and other sounds considered as noise.

A bone conduction microphone connected wirelessly to a smartphone using Bluetooth was used for dietary activities sound collection. An application software has been developed to perform data collection. The smartphone used was a Google Pixel 3, and the bone conduction microphone was a Motorola Finiti HZ800 Bluetooth Headset. The sound signal sampling from the microphone was 8kHz. After collection, data were transferred to a computer for labeling and analysis. Besides, since data were collected in a free environment, it was necessary to perform labeling afterwards.

Data from 16 men and women aged from 11 to 23 years old were collected. Subjects include children who are still in development to obtain a larger variety of data, such as different chewing patterns between children and adults. Informed consent for data collections was obtained from the subjects or their legal guarantor. As shown in Figure 1, which reproduces the data collection conditions, subjects put on one ear the bone conduction microphone that communicates via Bluetooth with the smartphone. Also, we shot a video to assist the afterward labeling task. Videos focused on the mouth and throat of the subjects. Data were collected in a natural meal environment and participants were required to have a usual meal as every day. 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. The meal content was also totally free, and participants ate whatever they wanted as usual in daily life, such various food types were mixed unpredictably during the same meal. Besides, the collected data time varied by cases collected from the start of the meal or the middle of the meal.

3.2 Segmentation and Labeling of audio data

When labeling audio data collected by the bone conduction microphone, video taken simultaneously with the recording of the audio data also used because it is difficult to label using only audio data. 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. Audio data were labeled and segmented manually using Praat [17], which is an audio analysis software that can associate labels to audio data sections. It takes about one hour to label for one-minute audio data. The labels were set according to the targeted five eating behaviors: "chewing" (C), "swallowing food" (S), "swallowing drink" (Drink), "talking" (T), and "other" (O). Each frame-size for labeling is different because a label was given for each eating activity. The video synchronized audio data were used as a reference. Labeling was performed by referencing the raw data and intensity of the audio signal and the synchronized video like in Figure 2. To

Fig. 1 Picture reproducing the data collection conditions

synchronize two recording data, characteristic sound like clap was made as a mark before starting the meal.

Table 1 sums-up the characteristics of the obtained dataset. Since our purpose is to tackle daily natural eating conditions, it is essential to collect data from an uncontrolled eating environment. Though scripting the dataset would have made the analysis more manageable, it would have created a bias on the data since the subject would have eaten, taking care of the script. Such, not all subjects have a similar amount of data, and the amount of data per class is imbalanced. Each subject has a different number of times of meals. Some subjects were collected data for several times meal, while others were collected from the middle of one meal. The number of data labeled as "chewing" is overwhelmingly imbalanced compared to other data, which were labeled "swallowing food", "swallowing drink", "talking", and "other". However, it is natural since, during a meal, one chew more than 10 times more than swallow food.

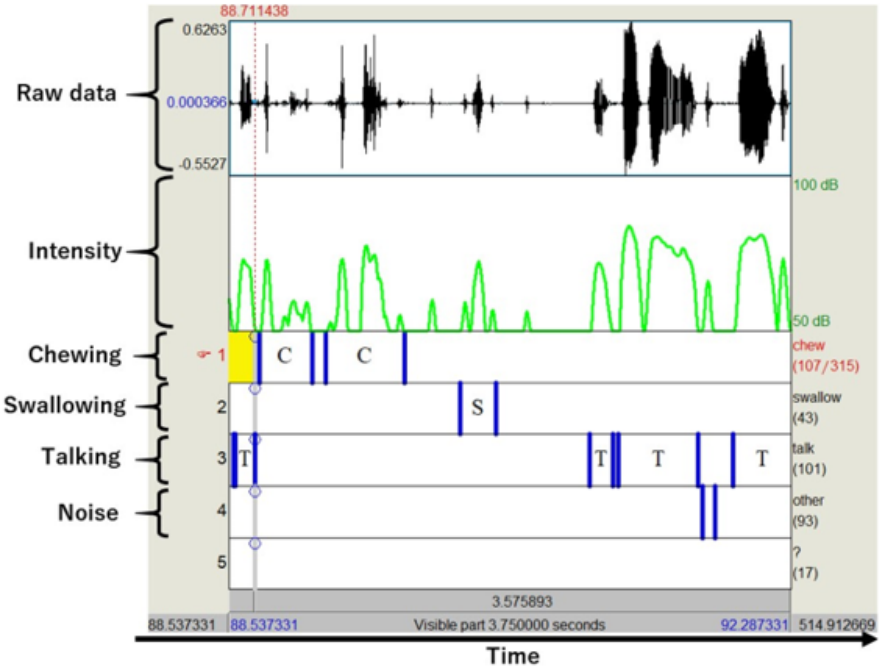

Fig. 2 Labeling with Praat

Table 1 The number of each labeled data

Subject	chewing	swallowing food	swallowing drink	talking	other
1	566	39	15	167	48
2	181	10	0	39	10
3	502	14	14	64	16
4	86	12	2	60	12
5	102	2	0	8	6
6	209	8	0	65	6
7	109	7	0	45	24
8	157	21	0	50	46
9	89	6	0	6	13
10	0	0	5	14	2
11	0	0	7	12	0
12	0	0	18	10	8
13	0	0	2	2	4
14	0	0	4	4	2
15	0	0	10	2	4
16	0	0	6	7	0
Total	2001	119	83	555	201

3.3 Feature extraction

In this research, features were extracted from the dataset labeled according to the previous section before operating machine learning for the classification. 75 features were extracted using Matlab (Mathworks, Inc.) to improve classification accuracy. Table 2 shows the extracted features.

Table 2 Extracted 75 features

Features category	Description	Number of features
Samples characteristics	number of samples number of peaks maximum magnitude zero-crossing count	4
Amplitude difference accumulation	sum of data points maximum value	2
Short term energy	sum of data points maximum value	2
Moving median value	zero-crossing count	1
Power Spectrum (PSD)	central frequency 1st and 2nd peak frequencies band total power density band average power density band peak frequency	23
Cross-correlation	raw data cross-correlation peak PSD cross-correlation peak	4
MFCCs	Mel-Frequency cepstrum Coefficients	39

As shown in the Table 2, four features were extracted from raw data sample characteristics: the time length of the sound section (number of samples), the number of peaks, the maximum magnitude, and the number of zero-crossing. Since the time length of speaking is longer and produces more amplitude peaks than that of chewing and swallowing. It also shows that the magnitude of the speaking signal is more prominent than other sounds, and the magnitude of swallowing is smaller than other sounds. These features are expected to be useful to differentiate speaking with other behaviors. Amplitude Difference Accumulation (ADA) was also used to characterize the audio signal, as it has been revealed by Zhang et al. to be useful to characterize chewing from other sounds[18]. In Formula 1, the signal to be processed shows as x, number of samples in each frame shows as N, and nth ADA.

$$ADAn = \sum_{m=(n-1)*N+1}^{n*N} |x(m) - x(m-1)| \tag{1}$$

Short Term Energy (STE) is also widely used in previous studies to characterize the audio signal of chewing[5]. In Formula 2, the signal to be processed shows as s, the shift of the number of samples shows as n, and the window function shows as

w(n).

$$e(n) = \sum_{m=-\infty}^{\infty} (s(m) * w(n-m))^2 \tag{2}$$

One other features were calculated by using the zero-crossing count after applying the moving median to the raw signal. It was used because the number of amplitude peaks of chewing, swallowing, and speaking varies. From the frequency domain, power spectrum density (PSD) of raw data was used to produce five features. PSD was calculated by the Welch method to obtain smoother variations between the frequencies. PSD is frequently used in the voice recognition field. Five features were extracted: the 1st and 2nd largest peak frequencies, the central frequency, the total power, and the average power. Besides, the later three values were additionally extracted for specific frequency bands to produce 18 more features. In total, 23 features were extracted from each sample PSD.

In order to improve the differentiation accuracy of chewing and swallowing, one sample of each behavior has been selected and used as a template sample of both the raw data and the PSD. The cross-correlation of the dataset samples' raw data and PSD with each behavior's template has been calculated and the maximum result extracted, producing four features. Finally, 39 features were extracted from Mel Frequency Cepstrum Coefficients (MFCC), which are used frequently in voice recognition system[19]. From the above, 75 features were extracted for model learning and classification.

3.4 Balancing of the unbalanced dataset

The dataset made by labeling, as shown in Table 1, the number of data other than chewing was few despite many chewing data. Therefore, first, a dataset that reduced the number of chewing data was created. Considering the number of speaking samples, which is the second-largest label, 500 chewing samples were randomly selected from the 2001 samples. The result is the first line of Table 3. However, the dataset is still unbalanced. Thus, Synthetic Minority Oversampling TEchnique (SMOTE) was used to resample the dataset[20]. SMOTE is a method proposed by Chawla that combines oversampling of the minority class and undersampling of the majority class. There exist several types of SMOTE. In this research, SVM-based SMOTE was used because it enables the generation of new samples on a wider area[21, 22]. Besides, SVM-SMOTE was used after dividing the original dataset into a training dataset and a test dataset, such only the training dataset was resampled. This avoids including test samples that would have not been real. This time, the dataset was divided into training data and test data by 4:1 and then only training data were resampled by using SVM-SMOTE. Since at each execution of SVM-SMOTE the resulting number of samples varies slightly, the second line of Table 3 represents one example of the numbers obtained after resampling.

Table 3 Example of the number of samples obtained after data selection and resampling

Label		chewing	swallowing food	swallowing drink	speaking	other
Number of data	selecting chewing data	500	119	82	554	200
	using SVMSMOTE (training data)	447	222	295	447	447

3.5 Evaluation of the learning accuracy of various classifiers

In this research, machine learning was used to build models that can classify chewing, swallowing food, swallowing drink, speaking, and other sounds from bone conduction audio data.

"Classification Learner App", which is the application in Matlab, was used to find out which classification model is best. This application can execute the automatic learning and find out the classification model with the best accuracy.

The "Classification Learner App" outputs the score of each model according to the specified validation method. In this research, 10-folds cross-validation was specified and the average accuracy for each method was output, as summed-up in Table 4.

Table 4 10-folds cross-validation result for the classification models evaluated

Classification models		Accuracy[%]
Decision Trees	Fine Tree	68.6
	Medium Tree	62.8
	Coarse Tree	54.6
Support Vector Machines (SVM)	Linear	73.1
	Fine Gaussian	62.8
	Medium Gaussian	85.9
	Coarse Gaussian	68.2
Nearest Neighbor Classifiers	Fine KNN	80.2
	Medium KNN	67.3
	Coarse KNN	54.0
	Cosine KNN	74.2
	Cubic KNN	65.1
	Weighted KNN	75.0
Ensemble Classifiers	Boosted Trees	68.5
	Bagged Trees	82.4

As shown in Table 4, medium Gaussian SVM model obtained the highest average accuracy. Therefore, SVM using Gaussian kernel (rbf kernel) was selected to classify eating behavior from bone conduction audio data collected in a natural meal environment.

4 Validation of Proposed Classification Model General Performances

This chapter reports and discussed the validation of the performance of the classification model. The test dataset is used to validate the classification accuracy of detailed eating activity from bone conduction audio data collected in a natural meal environment.

4.1 The classification model

The classification model used in this research is SVM with rbf kernel selected by the result of the "Classification Learning App" available in Matlab (Mathworks, inc.). Using SVM requires the scaling of feature values. In this research, each feature value was scaled such as the average is zero and the variance is one. Also, adjustment of the SVM classifier parameters was required to classify with higher accuracy. The parameters of rbf kernel SVM are the regularization parameter "C," and the inverse of the Gaussian kernel width "gamma." "C" and "gamma" 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[23]. In this research, the method called "grid-search" was used to adjust the parameters. Using "grid-search" a total of 36 combinations of "C" and "gamma" were evaluated by cross-validation. The results of "grid-search" show that, optimal parameters of SVM when using 75 features are 10 for "C" and 0.01 for "gamma."

4.2 Generalization performance results of the optimized model

Generalization performance results of SVM adjusted with the optimal parameters by using test data were evaluated. The general performance is the ability of a model to predict unknown data accurately. This section describes the SVM generalization performance results for four and five eating behaviors.

First, classification was performed for four labels: chewing, swallowing, speaking and other sounds. The label of swallowing includes both swallowing food and drink, as in previous research[16]. Second, classification was performed for five labels: chewing, swallowing food, swallowing drink, speaking, and other sounds.

Next, generalization performance results were produced for each number of labels in the case of optimized Gaussian kernel SVM using 75 features and using reduced 48 features. The reason why the features number was reduced from 75 to 48 is that 12th MFCC was used instead of the 39th MFCC. The 39th MFCC was used from the previous research. This research also proposed to use only the first 12 MFCC because it is used frequently in voice recognition.

The results were precision, recall, and F1 score for each class, according to the following definitions.

- The precision is the rate of how much the prediction was correct.
- The recall is the rate of correct prediction in each class.
- The F1 score is harmonic mean of precision and recall.

Also, this time, shuffle-split cross-validation with 10-folds was performed. As a result, a combination of training data and test data was prepared 10 times, resampling the training data each time, and generalization performance results were produced using the test dataset. For each result, the average was calculated.

4.2.1 Generalization performance results for 4 labels with 75 features

In this section, generalization performance results in the case of four labels and 75 features is stated. Table 5 shows the average value of the 10 shuffle-splits.

Table 5 Average classification performance results

	precision	recall	F1-score
chewing	0.73	0.78	0.75
swallowing	0.47	0.41	0.43
speaking	0.90	0.90	0.90
other	0.52	0.50	0.51

From these results, the value of F1 score of chewing and speaking were over 75%. On the other hand, the value of F1 score of swallowing and labeled other were around 50%. The overall classification accuracy obtained from the test data and the predicted labels was 73% on average, the maximum was 75%, and the minimum was 71%.

4.2.2 Generalization performance results for 4 labels with 48 features

In this section, generalization performance results in the case of four labels and 48 features is stated. Table 6 shows the average value of the 10 shuffle-splits.

Table 6 Average classification performance results

	precision	recall	F1-score
chewing	0.78	0.78	0.78
swallowing	0.49	0.49	0.48
speaking	0.90	0.88	0.89
other	0.48	0.52	0.50

From these results, the value of F1 score of chewing and speaking were over 75%. Also, the value of F1 score of swallowing and labeled other were around 50%. The

overall classification accuracy obtained from the test data and the predicted labels was 74% on average, the maximum was 81%, and the minimum was 68%.

4.2.3 Generalization performance results for 5 labels with 75 features

In this section, generalization performance results in the case of five labels and 75 features is stated. Table 7 shows the average value of the 10 shuffle-splits.

Table 7 Average classification performance results

	precision	recall	F1-score
chewing	0.71	0.84	0.77
swallowing food	0.31	0.26	0.28
swallowing drink	0.28	0.16	0.19
speaking	0.90	0.91	0.91
other	0.58	0.46	0.51

From these results, the value of F1 score of chewing and speaking was over 75%. On the other hand, the value of F1 score of labeled other was 51%. Also, the value of F1 score of swallowing food and swallowing drink was under 30%. The overall classification accuracy obtained from the test data and the predicted labels was 73% on average, the maximum was 77%, and the minimum was 69%.

4.2.4 Generalization performance results for 5 labels with 48 features

In this section, generalization performance results in the case of five labels and 48 features is stated. Table 8 shows the average value of the 10 shuffle-splits.

Table 8 Average classification performance results

	precision	recall	F1-score
chewing	0.74	0.75	0.75
swallowing food	0.27	0.28	0.27
swallowing drink	0.32	0.27	0.29
speaking	0.92	0.88	0.90
other	0.51	0.56	0.54

From these results, the value of F1 score of chewing and speaking was 75% and 90%. On the other hand, the value of F1 score of labeled other was 54%. Also, the value of F1 score of swallowing food and swallowing drink was 27% and 29%. The overall classification accuracy obtained from the test data and the predicted labels was 71% on average, the maximum was 75%, and the minimum was 66%.

4.3 An example of the distribution of predicted result

In this section, an example of the distribution of the actual labels in test data classification is described, based on the number of predicted labels. The results for one of the tests are shown because the test was performed 10 times. The number was counted with the correct label of test data for each predicted labels.

Table 9 shows the result with five labels and 75 features.

Table 9 One example result of classification with 5labels and 75 features

		True label				
		chewing	food swallowing	drink swallowing	speaking	other
Predict label	chewing	86	13	7	5	17
	food swallowing	3	3	3	0	3
	drink swallowing	2	5	4	0	1
	speaking	0	1	2	107	2
	other	7	1	2	1	16

5 Discussion

5.1 Discussion about the generalization performance results

Considering the generalization performance results for all case, the F1 score of chewing and speaking resulted in at least 75% accuracy. In particular, speaking gained a good result with over 85% accuracy. However, the result of swallowing and other sounds, especially swallowing food and swallowing drink was poor. Also, it was found that actual swallowing label were often predicted as chewing. Also, most of data that were predicted swallowing drink label and it is not actual label, were often predicted swallowing food and chewing. The actual food and swallowing drink label were predicted mostly chewing.

Considering the differences in the results depending on the number of labels there was no big difference. The F1 score of labels related to swallowing behavior decreased when the number increased from four to five. The swallowing label was divided into two labels according to the differences of the content of swallowing, but each F1 score was lower than when classifying as swallowing.

From the above discussion, it was found that there is no significant difference in classification accuracy between using 12th MFCC and 39th MFCC as one set of the features when classifying from sound signal meal behaviors. Therefore, it was shown

that the same level classification accuracy can be maintained even with a smaller number of features by using 12 MFCC instead of 39.

Further, the result of this method that resample after dividing into training data and test data showed low accuracy compared with previous work[16]. Therefore, resampling before dividing caused model to overfit.

5.2 Discussion about the dataset

In this research, labeling was performed among five labels to be able to classify more detailed eating behaviors. It resulted in low classification accuracy. The number of swallowing, especially swallowing drink, was significantly less than that of chewing in these collected data. Therefore, more data samples of swallowing food and drink sound are required to improve classification accuracy. Also, it is necessary to find features that express better the differences between these three behavior of chewing, swallowing food and swallowing drink. One can expect that adding new features may improve the differentiation between these three behaviors.

Besides, it was very difficult to perform the labeling of the swallowing label. Although labeling was performed using the synchronized video as a reference, especially labeling swallowing food was difficult to judge. Some of the reasons were for example, unnoticeable movement of women's throat or tableware hiding the throat. Therefore, what is labeled as swallowing may not be swallowing actually. Also, there may be bias because only one person labeled the data used in this research. It is necessary to perform more correctly and less bias labeling, such as labeling by several people and adopting only similar ones.

6 Conclusion and Future Work

6.1 Conclusion

This research proposed a classification method of eating behavior from dietary sound collected in a natural meal environment.

Two models using SVM with rbf kernel were built and optimized for the classification of four eating behaviors (chewing, swallowing, speaking and other sounds) and five eating behaviors (chewing, swallowing food, swallowing drink, speaking and other sounds). The generalization performance for each model was compared with when using 39 MFCCs and when using 12 MFCCs in addition to other common features. As a result, it was demonstrated that the difference in the number of features did not affect the classification performance of unknown samples. The F1 score of chewing and speaking resulted with high accuracy. Therefore, detection of chewing and speaking is possible by using this method. Also, the F1 score of swallowing when using four labels, and that of swallowing food and swallowing drink when using five

labels was very low. From there, it was found that further devices of used features are important.

6.2 Future work

As a prospect, using the proposed classification method to judge detailed eating behaviors in real-time by using a bone conduction microphone and a smartphone is planed. For that purpose, it is necessary to search for a more accurate classification method. This includes reviewing the labeling quality, optimizing the resampling method, and adding or changing used features. Also, it requires to design a system that can automatically extract sections of audio data that are related to the dietary activity such as chewing, swallowing and speaking.

Besides, five types of eating behavior have been considered in this study, but the eating behaviors can be quantified with more details such as the amount of ingested liquid, or the texture of chewed food (hard, soft, crispy, crunchy, etc.). Also, only the data from young subjects have been collected. More data is needed to classify regardless of age or gender. In particular, data of swallowing food and swallowing drink have fewer samples, so collecting more samples of these behaviors is essential. Finally, although participants were instructed to eat freely, there are may be other dietary sounds that have not been recorded such as soup or noodles slipping in the mouth.

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. Yuichi Ando, Nobuhiro Hanada, and Shigetaka Yanagisawa. (2008). Does "eating slowly" lead to prevent obesity ?. *Health Science and Health Care, 8(2), 51–63.*
4. Noriko Kishida., and Yoshie Kamimura. (1993). Relationship of conversation during meal and health and dietary life of school children. *The Japanese Journal of Nutrition and Dietetics, 51(1), 23–30.*
5. Hideto Mitsui, Joe Ohara, Anna Yokokubo, and Guillaume Lopez. (2018). Method to Improve Real-time Chewing and Speaking Detection Accuracy from Bone-conduction Sound. *Multimedia, Distributed, Cooperative, and Mobile Symposium 2018, 2018: 562–566.*
6. Rui Zhang, Severin Bernhart, and Oliver Amft. (2016). Diet Eyeglasses: Recognising Food Chewing Using EMG and Smart Eyeglasses. *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 7–12.*
7. Keum San Chun, Sarnab Bhattacharya, and Edison Thomaz. (2018). 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.*
8. Shuangquan Wang, Gang Zhou, Yongsen Ma, Lisha Hu, Zhenyu Chen, Yiqiang Chen, Hongyang Zhao, and Woosub Jung. (2018). Eating detection and chews counting through sensing mastication muscle contraction. *Smart Health, 9, 179–191.*

9. Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. (2020). FitByte: Automatic Diet Monitoring in Unconstrained Situations Using Multimodal Sensing on Eyeglasses. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12.*
10. Oliver Amft, Mathias Stäger, Paul Lukowicz, and Gerhard Tröster. (2005). Analysis of chewing sounds for dietary monitoring. *International Conference on Ubiquitous Computing, 56-72.*
11. Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, et al. (2018). Auracle: Detecting Eating Episodes with an Ear-mounted Sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(3), 92.*
12. 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), 158–166.*
13. Hao Zhang, Guillaume Lopez, Ran Tao, Masaki Shuzo, Jean-Jacques Delaunay, and Ichiro Yamada. (2012). Food Texture Estimation from Chewing Sound Analysis. *HEALTHINF, 213–219.*
14. Mark Mirtchouk, Drew Lustig, Alexandra Smith, Ivan Ching, Min Zheng, and Samantha Kleinberg. (2017). Recognizing Eating from Body-Worn Sensors: Combining Free-living and Laboratory Data. *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Vol. 1, No. 3, Article 85.*
15. Takumi Kondo, Hidekazu Shiro, Anna Yokokubo, and Guillaume Lopez. (2019). Optimized Classification Model for Efficient Recognition of Meal-Related Activities in Daily Life Meal Environment. *Proc. of the 2019 joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 146–151.*
16. Takumi Kondo, Haruka Kamachi, Shun Ishii, Anna Yokokubo, and Guillaume Lopez. (2019). Robust classification of eating sound collected in natural meal environment. *Proc. of the 2019 ACM International joint Conference on Pervasive and Ubiquitous Computing and 2019 ACM International Symposium on Wearable Computers, 105–108.*
17. Praat. http://www.fon.hum.uva.nl/praat/
18. Hao Zhang, Guillaume Lopez, Masaki Shuzo, Jean-Jacques Delaunay, Ichiro Yamada. (2012). Mastication Counting Method Robust to Food Type and Individual. *HEALTHINF, 374–377.*
19. Jumpei Ando, Takato Saito, Satoshi Kawasaki, Masaji Katagiri, Daizo Ikeda, Hiroshi Mineno, and Masafumi Nishimura. (2017). Conversation and Eating Behavior Recognition by Leveraging Throat Sound. *Multimedia, Distributed, Cooperative, and Mobile Symposium 2017, 2017,116–123.*
20. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16, 321–357.*
21. Lianbao Yang, Ping Li, Rui Xue, Xiaoning Ma, Xinqin Li, and Zhe Wang. (2018). Intelligent classification model for railway signal equipment fault based on SMOTE and ensemble learning. *IOP Conference Series: Materials Science and Engineering, 383(1), 012042.*
22. imblearn.over_sampling.SVMSMOTE. https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SVMSMOTE.html
23. Andress C. Muller and Sarah Guido. (2017). Introduction to Machine Learning with Python. *O'Reily.*