Bone Conduction Eating Activity Detection based on YAMNet Transfer Learning and LSTM Networks

Wei Chen[1][a], Haruka Kamachi[1][b], Anna Yokokubo[2][c] and Guillaume Lopez[2][d]
[1] *Graduate School of Science and Engineering, Aoyama Gakuin University, Sagamihara, Japan*
[2] *Department of Integrated Information Technology, Aoyama Gakuin University, Sagamihara, Japan*

Keywords: Bone Conduction, Transfer Learning, Long-short Term Memory (LSTM), Eating Behavioural Activity.

Abstract: The trivial eating behaviors affect our health and sometimes lead to obesity and other health problems. We propose an automatic human eating behavior estimation system , which performs real-time inferences using a sound event detection (SED) deep learning model. In addition, We customized YAMNet, a pre-trained deep neural network by 521 audio event classes based on Mobilenet v1 depthwise-separable convolution architecture from Tensorflow. We used transfer learning shaped YAMNet as a feature extractor for acoustic signals and applied an LSTM network as a classification model that can effectively handle time-series environmental acoustic signal. Dietary events including chewing, swallowing, talking, and other (silence and noises), were collected on 14 subjects. The classification results show that our proposed method can validly perform semantic analysis of acoustic signals of eating behavior. The overall accuracy and overall F1 scores were both 93.3% in frame level, respectively. The classifier established in this study provided a foundation for preventing premature eating and a healthier eating behavior monitoring system.

1 INTRODUCTION

In modern life, the data of all human habits are being digitized for a healthier lifestyle. Automatic detection of dietary habits is one of the challenges of human habits digitization. This paper explored a method for automated eating activity using a commercially available bone conduction microphone. Compared to conventional methods for automatic detection in eating activity analysis-related works, this paper focuses on improving the accuracy rate of each independent eating activity identification, with the basic premise of using acoustic signals from the natural environment.

According to the 2016 global obesity population distribution by WHO, Approximately 39% of the world's population is overweight, of which 13% is obese (NCD Risk Factor Collaboration, 2016). In addition, surveys of obese people have shown that many of them are ”fast eaters” who chew less and eat for a shorter time (Yamaji et al., 2018). To prevent obesity, automatic detection of eating behavior using wearable devices has been progressing over the past decade (Selamat and Ali, 2020). Wearable sensors that have been proposed for automatic detection of eating behavior include in-ear microphones (Amft, 2010; Shuzo et al., 2010), neck-worn sensors(Chun et al., 2018), strain sensors (Yang et al., 2019), electromyography sensors (Huang et al., 2017), and wrist-worn sensors (Shen et al., 2016).

As the most used method, acoustic sensing is one of the earliest modalities studied, with advantages such as ease of wear and precise identification of chewing. Being able to strike a balance between high-quality signal acquisition and user comfort is the main challenge of acoustic eating activity sensing. Kamachi et al. proposed a classification method of eating behavior by capturing both the chewing sound based on bone conduction microphone to capture both chewing and swallowing sound (Kamachi et al., 2021). Päßler et al. used their proposed design to perform analysis such as analyzing acoustical signal energies and chewing detection based on magnitude squared coherence function (MSC) (Päßler and Fischer, 2011) . The fact that teeth produce vibrations when they tap, slide, or grind against each other, these vibrations travel through the jaw and skull bones as surface vibrations can easily reach the outer ear (Prakash et al., 2020), leads to a high accuracy

[a] https://orcid.org/0000-0001-7951-137X
[b] https://orcid.org/0000-0002-9269-1026
[c] https://orcid.org/0000-0003-2657-4961
[d] https://orcid.org/0000-0002-9144-3688

Chen, W., Kamachi, H., Yokokubo, A. and Lopez, G.
Bone Conduction Eating Activity Detection based on YAMNet Transfer Learning and LSTM Networks.
DOI: 10.5220/0010903700003123
In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - Volume 4: BIOSIGNALS*, pages 74-84
ISBN: 978-989-758-552-4; ISSN: 2184-4305

rate of chewing identification. However, there are still some issues to be left, such as difficulty in recognizing swallowing and susceptibility to background noise (Kamachi et al., 2020).

Most of the eating activity detection methods in the previous papers are based on manually extracted features from speech signals and well-researched classification algorithms such as support vector machines (Zhang et al., 2011; Nkurikiyeyezu et al., 2021). In recent years, with the advancement of deep neural networks, classification techniques using deep learning such as Convolutional neural networks (CNN) and Recurrent neural networks (RNN) are also suitable for acoustic signals (Bae et al., 2016; Xu et al., 2018). An emerging approach is to apply transfer learning to the recognition of acoustic signals (Ntalampiras et al., 2021).

This work uses acoustic signals recorded from bone conduction microphones as input and trains a model that combines transfer learning and deep learning to incorporate them into the automatic eating activity detection task compatible with natural environments. The experimental results show that our model significantly improves effectiveness compared to existing state-of-the-art approaches, which is very promising.

This paper is organized as follows. Section 2 presents the processing pipeline focusing on the architecture, detail of the datasets, and the primary method used in this study. In Section 3, we describe the classification method, evaluation methods, and experimental results. In section 4, we discuss our proposed method compared with the previous ones and present future work. Finally, we conclude this paper in Section 5.

This section describes the sample data utilized in this study and the methods used to process the audio signal data, the feature extractor, and the classifier model parameters.

1.1 Eating Behavior Signal Data Collection

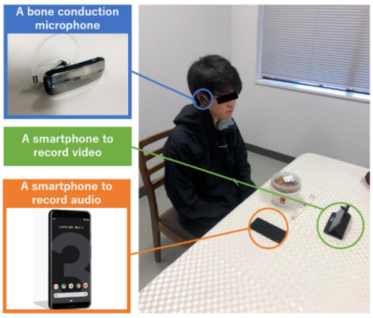
A bone conduction microphone
A smartphone to record video
A smartphone to record audio

Figure 1: Data collection environment and devices.

2 MATERIAL AND METHOD

To evaluate the segmentation method for detection eating activities using bone conduction sound, we collected meal sound data in a natural meal environment. We used a bone conduction microphone connected wirelessly to a Smartphone using Bluetooth protocol for dietary activities sound 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 44100 Hz. After collection, we transferred data to a computer for labeling and analysis at 16000 Hz. Besides, it was necessary to perform labeling afterward since data collection in a free environment.

We collected data from 14 participants. All of the young subjects were between the ages of 11 to 32 years. As shown in Figure 1, which reproduces the data collection conditions, subjects put the bone conduction microphone on one ear. Also, we shot a video focused on the mouth and throat of the subjects in order to assist the afterward labeling task of audio sections corresponding to chewing, swallowing, talking, and other sounds (like noise). The participants will be asked to say a certain word at the beginning of the experiment, which will be used to synchronize the video and audio data.

To provide a natural environment, we collected data in general daily life, such as big surrounding sounds and eating with a conversation. Participants were required to have a usual meal as every day's meal. For example, in a dining room, a standard household table with other family members, and at the university cafeteria with friends, we assume that represents different noisy conditions. 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. Also, the collected data time varied by cases collected from the meal's start or the middle of the meal. Besides, we collected additional swallowing sound data because the few swallowing data compared to other classes have been pointed out in our previous works. The swallow audio data from 8 men and women aged from 22 to 42, who were required to have a couple of drinks, were collected the same as above.

2.1 Architecture of the Proposed Method

Our proposed architecture is shown in Figure 2. The first step is to label the acquired data so that it may

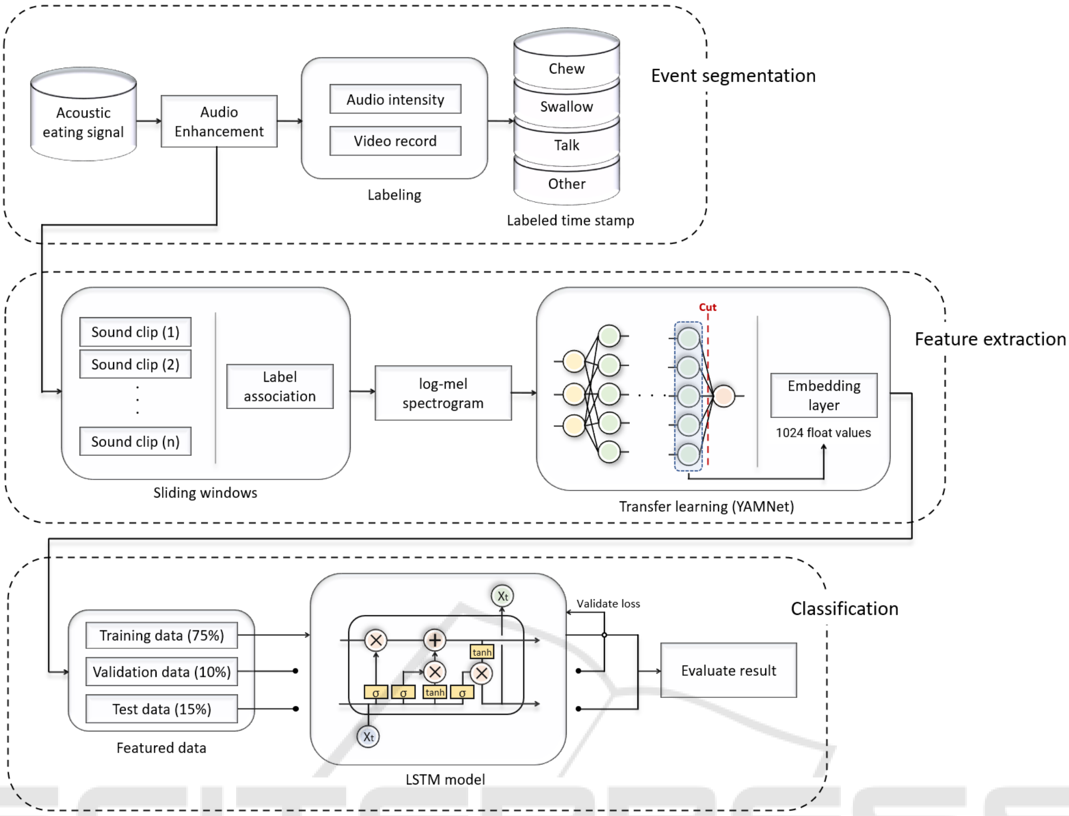

Figure 2: Block Diagram overview in this study.

be used as a reference for training data to predict eating behavior. The acoustic eating behavior data taken from a bone conduction microphone was manually selected to include chewing, swallowing, speech, and other sound events including noises applicable segments. In the next stage, the labeled sound episodes are prepossessed by window segmentation, acoustic enhancement and the features of the data are extracted from transfer learning using YAMNet (Tensorflow, 2020). Finally The resulting embedding layer is then classified using a deep learning network Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) to evaluate the classifier.

2.2 Sound Episode Segmentation

In order to segment the eating behavior to the specific number of bites, the acoustic signal of a single bite has to be accurately labeled. For the labeling process, we used speech intensity to more easily segment the speech signal (Clark et al., 2014). As the purpose of this study is to estimate human eating behaviour automatically and possible to classify the audio signal in real time, We define eating behavior as four steps: chewing, swallow, talk and other. Chewing refers to the vertical opening and closing of the top and bottom teeth during a single chew. Swallowing refers to the swallowing of food in one sitting. It is also considered a single swallow if the subject drinks. Talk refers to what the subject is vocalizing. Other refers to events that are not all of the above event, such as when there is no sound or when there is noise. The number of each sound event we labeled is shown in Table 1. As you can see the data is not balanced well due to the number of swallow is low. Because of this is also in agreement with our philosophy of chewing as much as possible in a single swallow during eating behavior.

We define one acoustic eating episode label by comparing the raw acoustic data with the sound intensity and referring to the beginning and end points of the waves that are visually obvious as shown in Figure 3. The amplitude of swallow episode in the acoustic data acquired from the bone conduction microphone tends to be smaller than that of chewing episode. Therefore, the acoustic data was recorded and synchronized with video to label specific swallows.

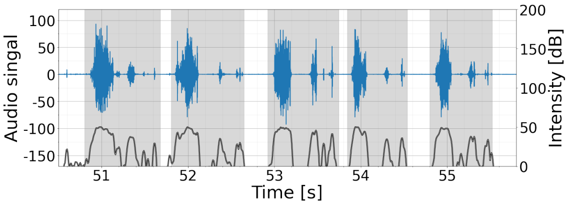

Figure 3: Part of continuous chewing data labeling scene.

Table 1: The number of datasets collected.

Sound episode categories	Data number
Chewing data	3395
Swallow data	334
Talk data	491

2.3 Feature Extraction and Signal Processing

The most commonly used features in acoustic signal processing are Mel-Frequency Cepstral Coefficients (MFCC) (Ittichaichareon et al., 2012). As a mechanism of MFCC, the magnitude spectra projected on the reduced frequency bands are transformed into logarithmic magnitudes, which are then approximately whitened and compressed by discrete cosine transform (DCT). In contrast, DCT removes information and destroys spatial relations in deep learning models, most audio signal processing with deep learning methods use log-mel spectrograms to perform feature extraction (Purwins et al., 2019) (Lee et al., 2017) (Zheng and Yan, 2019).

Also, feature extraction methods based on log-mel frequencies using transition learning have emerged. In the evaluation of the Non-Semantic Speech benchmark (NOSS), which assess the general usefulness of speech representations on "non-semantic" tasks, shows that the correctness of the middle layer outputs of YAMNet and TRILL all reach a good accuracy rate (Shor et al., 2020). In the same acoustic signal classification domain, the YAMNet model is treated as a feature extractor by outputting an intermediate layer embedding using transfer learning for COVID-19 cough classification (Elizalde and Tompkins, 2021).

2.3.1 Pre-emphasis

Pre-emphasis is a widely used method in audio signal processing which has the effect of emphasizing the wide-area components of the audio waveform (Dong et al., 2020). In this study pre-emphasis is performed on the raw signal before it is processed to compensate for the frequencies in the high frequency portion of the acoustic signal. The following filters are used shown as equation (1):

$$Au'(n) = Au(n) - \alpha Au(n-1) \tag{1}$$

where, Au and Au' are the raw audio signal before and after the pre-emphasis operation; n is the index of each sample in raw audio signal; α is the parameter which was normally assigned a value in the range of $[0.9, 1]$.

2.3.2 Framing and Windowing

After pre-emphasis, the raw audio signal is split into slide windows to generate a log-mel spectrogram. statistics of the collected data showed in Table 2. The average time of one chewing is around 315ms. Also There are very short sound episodes in the labeled data. If the label time is larger than the window size, the labeling will be inaccurate, so in order to effectively classify more sound episode clips, We define the window size as 250ms which is below the mean time of chewing data event. and the hop size of each window as 93ms, which is less than one-half of window size and greater than one-third of window size.

Table 2: Statistics of each label timing.

Categories	$Mean_{(ms)}$	$Min_{(ms)}$	$Max_{(ms)}$	STD
Chewing	314	28	1042	0.1572
Swallow	405	54	1541	0.2353
Talk	859	92	5067	0.6714

2.3.3 Log-mel Spectrogram

Not only YAMNet, but also many other acoustic deep learning methods use mel spectrograms as input preprocessing form for audio signals (Zeng et al., 2019). According to the YAMNet summary, we generate the mel-scaled spectrograms with a triangular filterbank of 64 log-energies. The relationship between the Mel spectrum and the frequency is shown in equation (2):

$$f_{mel} = 2595 \cdot lg(1 + \frac{f}{700Hz}) \tag{2}$$

where, f_{mel} is the mel frequency; f is the linear frequency.

Then use the short-time Fourier transform (STFT) to find out the frequency of shorter intervals. We defined the window size of STFT to be 25ms and the hop size of STFT to be 10ms. The progress feeding the signal into the filterbanks to get the $H_m(k)$ is shown in equation (3):

$$H_m(k) = \begin{cases} \dfrac{k - f(m-1)}{f(m) - f(m-1)}, f(m-1) \leq k \leq f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, f(m) \leq k \leq f(m+1) \\ 0, \mathit{others} \end{cases} \tag{3}$$

where, $f(m)$ is the m^{th} filter's centre frequency; $H_m(k)$ is the returned filterbank as matrix.

The result of multiplying filterbanks by STFT, which is the raw audio signal processed to energy spectrum, is shown in equation (4):

$$LogMelSpec(m) = \sum_{k=f(m-1)}^{f(m+1)} log(H_m(k) \cdot |X(k)^2|) \tag{4}$$

where, $|X(k)^2|$ is the energy spectrum is the point of k^{th} energy; m is the filterbanks and k is the point of FFTs.

2.3.4 YAMNet Embedding

YAMNet is a pre-trained deep network which predicts 521 audio event classes based on the AudioSet-YouTube corpus (Gemmeke et al., 2017). Employing the Mobilenet_v1 (Depthwise-separable convolution) architecture (Howard et al., 2017). The audio set for training the YAMNet model contains more than 632 audio events sampled from a 10-second clip of a YouTube video that has been played more than 1000 times. Due to the properties of deep learning and YAMNet, a feature extraction layer is built into the model. Therefore, the log-mel spectrogram of the speech signal directly becomes the input for Mobilenet_v1.

The network operators on input mel spectrogram of size $(48, 32, 32)$ as we get in previous section. The structure of YAMNet is shown in Table 3. Inputs signal is processed by an 1-D convolution layer, which the kernel size of 3×3. Then pass the value through the number of filter 64-1024. The global average pooling (AP) layer is in the next to prevent potential over-fitting by reducing the total number of parameters of the model. At last, the network comes with two fully connected (FC) layers of size 1024 and 64. With the last list of value and a softmax layer to compute one of 521 result determine which sound episode this log-mel spectrogram belong. The most important point of YAMNet is the last second fully-connected layer which have 1024 values. We customize the YAMNet network to contain with the network structure until last second fully-connected layer. We use YAMNet to output a 1024 values of embedding layer as feature extractor and treating YAMNet as a transfer learning method. The advantage of this method is that it has enough acoustic features even when the number of data for this signal classification problem is not so large as to overflow, and provides great trade-off between performance and computational cost.

Table 3: YAMNet body architecture.

Type	Filter shape	Input size
$Conv_1$	$3 \times 3 \times 3$	$48 \times 32 \times 32$
$Conv_2dw$	$3 \times 3 \times 3$ dw	$48 \times 32 \times 32$
$Conv_2pw$	$1 \times 1 \times 32 \times 64$	$48 \times 32 \times 64$
$Conv_3dw$	$3 \times 3 \times 64$ dw	$24 \times 16 \times 64$
$Conv_3pw$	$3 \times 3 \times 128$	$24 \times 16 \times 128$
$Conv_4dw$	$3 \times 3 \times 128$ dw	$24 \times 16 \times 128$
$Conv_4pw$	$1 \times 1 \times 128 \times 128$	$24 \times 16 \times 128$
$Conv_5dw$	$3 \times 3 \times 128$ dw	$12 \times 8 \times 128$
$Conv_5pw$	$1 \times 1 \times 128 \times 256$	$12 \times 8 \times 256$
$Conv_6dw$	$3 \times 3 \times 256$ dw	$12 \times 8 \times 256$
$Conv_6pw$	$1 \times 1 \times 256 \times 256$	$12 \times 8 \times 256$
$Conv_7dw$	$3 \times 3 \times 256dw$	$6 \times 4 \times 512$
$Conv_7pw$	$1 \times 1 \times 256 \times 512$	$6 \times 4 \times 512$
$Conv_8dw$ $-Conv_{12}dw$	$3 \times 3 \times 512$	$6 \times 4 \times 512$
$Conv_{13}dw$	$3 \times 3 \times 512$	$3 \times 2 \times 1024$
$Conv_{13}pw$	$1 \times 1 \times 512 \times 1024$	$3 \times 2 \times 1024$
$Conv_{14}dw$	$3 \times 3 \times 1024$ dw	$3 \times 2 \times 1024$
$Conv_{14}pw$	$1 \times 1 \times 1024 \times 1024$	$3 \times 2 \times 1024$
$AP\&Pool$ 3×2	$1 \times 1 \times 1024$	
FC	1024×512	$1 \times 1 \times 512$
$Softmax$	Classifier	$1 \times 1 \times 512$

2.4 Classifier Training and Evaluating

2.4.1 LSTM Model

Recurrent Neural Networks (RNNs) are very effective for analyzing sequences of text, acoustic signals, and video (Zhang and Man, 1998). The input signal can be persistently held by looping in the network. Basically, the main feature of RNN is that it remembers the previous state and uses that information to determine the next state. Therefore, models using RNN networks are very suitable for analyzing time series. However, the gradient of a traditional RNN depends not only on the present error, but also on the past error, so the retro-propagated gradients tend to grow enormously or fade over time.

LSTM is a type of RNN network that has a built-in function to determine which information to store and which to delete, and is a model that can store long-term dependencies without accumulating errors(Navarro et al., 2020). The LSTM module has

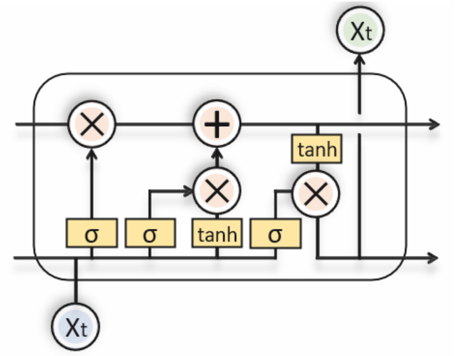

Figure 4: General structure of an Long Short-Term Memory neural networks (LSTM).

three internal gates, termed input, forgotten and output shown in Figure 4. The input gate controls when any new information will be put into memory. Forgotten gates allow the cell state to identify important and unwanted data when a piece of information is forgotten, leaving space for new data. The output gate is used to control the result of the memory stored in the cell state. The cell state has a weighting optimization mechanism and controls each gate based on the output error of the network. The cell state has a weighting optimization mechanism that controls each gate based on the output error of the network and sends the prediction to the next LSTM module.

2.4.2 Evaluation of Classifier

We use recall, specificity, precision, accuracy, and F1-score to evaluate the performance of the classifier. To calculate those values there are four types of possible results for the classification task. If the sample true label is positive and it is classified as positive is counted as a true positive (TP). If the sample is positive and it is classified as negative is counted as a false negative (FN). If a sample is negative and is classified as negative or positive, it is considered a true negative (TN) or false positive (FP), respectively. Based on them the result of recall, specificity, precision, accuracy, and F1-score is defined by the following equation:

$$Recall_{class} = \frac{TP_{class}}{N_{class}} \tag{5}$$

$$Specificity = \frac{TN}{FP+TN} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$F1_{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

3 RESULT

3.1 YAMNet Feature Extraction

The acoustic dietary data collected in this study was segmented and feature extraction was performed as described in Section 2.4, as shown in Figure 5. All wav files of recorded audio signals are resampled to 1.6khz, Pre-emphasised then cut into sliding windows of window size 250ms and hop size 93ms, which window size has 4000 sample points and hop size has 1500 sample points. The slide window is further applied to splitting the signal into short frames, applied STFT to generate a spectrogram. Finally applied to a 64 log-energies mel filterbank to output log-mel spectrogram prepared for feeding the YAMNet transfer learning.

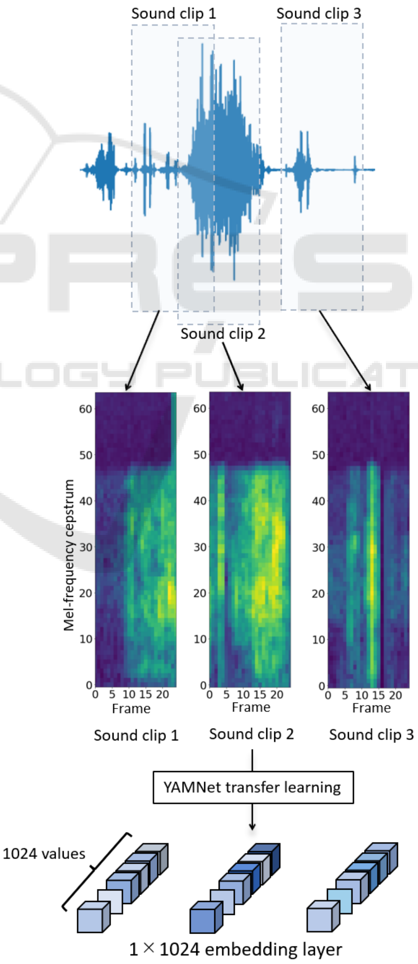

Figure 5: Sound clip splitting and generating the embedding layer as feature extractor using YAMNet.

3.2 LSTM Model Implementation

The feature extracted sound clips are divided into train data, validation data and test data. The number of data used for classification is shown in the following table 4. We set the ratio of train data, validation data and test data to be 75%-10%-15% (Vasudevan et al., 2020). As an input to the model, three sound clips are stored as a single continuous clip in order to increase the storage area in the time domain of the LSTM gate, and the overlapping of the three clips connected back and forth is performed. From the shape of our input data, we set the batch size, time steps, and feature to 16, 3, and 1024, respectively.

Table 4: Number of the sound clip for each label.

Category	Train	Validation	Test	Total
Chewing	8283	1155	1672	11304
Swallow	874	179	158	1129
Talk	3357	470	724	4504
Other	20913	2676	4134	27658
Total	33427	4480	6688	44595

The parameters of each layer of the LSTM model were set as follows and shown as table 5. The number of the hidden layers and the iterations for each is set to 128 and 256 (Altché and de La Fortelle, 2017). To prevent overtraining of the LSTM model, we set up a dropout layer with a rate of 0.25%, the dropout rate of 0.25% can effectively prevent memory loss in the model (Semeniuta et al., 2016). Finally, the Softmax layer outputs a vector with 4 elements to produce the results of the LSTM model, where the probability of each of the vectors is in the order of other data, chewing data, swallow data, and talk data. When the LSTM model complies, we fine-tune the model using the Adam optimizer. Adam is an optimization algorithm that can use instead of the classical stochastic gradient descent procedure to update network weights iteratively in training data (Kingma and Ba, 2014). The learning rate is set to 0.001, and the sparse categorical cross entropy is selected as the loss function. Use sparse categorical cross entropy when your classes are mutually exclusive such as when each sample belongs exactly to one class (Totakura et al., 2020) and expressed in Equation (10). Where, N presents the number of categories which is 4; y_i and $log\hat{y_i}$ respectively represents the label value and its log probability. The Model Loss on training and validation datasets is shown in Figure 6. It can be observed that there was no gradient explosion appeared from the plotted data. Train loss drops below 0.05 from 80 epochs and stays at about 0.02 from 200 epochs. Validation loss floats and stabilizes at about 0.5 from 120 epochs. We implemented 200 epochs, where the loss is stable, as a parameter during model training.

$$Loss = -\sum_{i=1}^{N} y_i \cdot log\hat{y_i} \qquad (10)$$

Table 5: LSTM body architecture for classification.

Type	Filter shape	Input size
$LSTM1$	kernel size 256	$1 \times 3 \times 256$
$LSTM2$	kernel size 128	1×128
$Dropout$	drop rate 0.25	1×128
FC	128×4	1×128
$Softmax$	Classifier	1×4

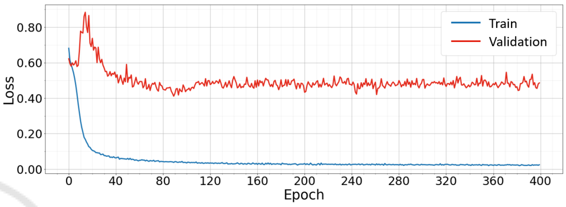

Figure 6: The model loss on training and validation datasets.

3.3 Classification Performances

As section 2.5.2 described the evaluation of classifier, a confusion matrix is to evaluate the performance of the classifier. The 6688 sound clip from the test datasets described in section 3.1 is used for evaluating, which shown as Figure 7.

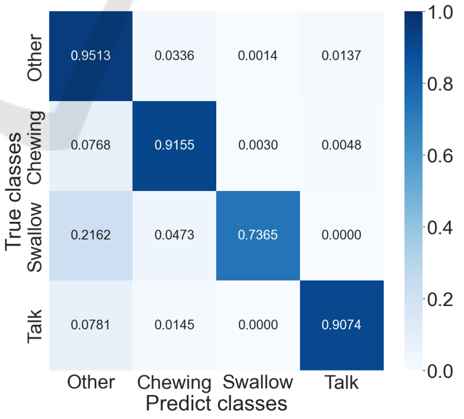

Figure 7: Confusion matrix of YAMNet as feature extractor.

According to the results of the confusion matrix, the sound clips of Chewing, Swallow, Talk, and Other, whose features were extracted by YAMNet, retain a significant percentage of accuracy. The accuracy rate was 91.54%, 73.64%, 90.73%, 95.13% respectively. The weighted average reaches 93.30%. And the overall F1-score reaches 93.28%. To understand the per-

Table 6: Evaluation result of the trained LSTM classifier each categories.

Evaluation metric	Chewing	Swallow	Talk	Other	Average	Weighted Average
Recall	91.55%	73.65%	90.74%	95.13%	87.77%	93.3%
Specificity	96.86%	99.83%	98.91%	91.46%	96.77%	93.78%
Precision	90.74%	90.83%	90.61%	94.86%	91.76%	93.29%
Accuracy	91.54%	73.64%	90.73%	95.13%	87.76%	93.30%
F1-Score	91.14%	81.34%	90.67%	94.99%	89.54%	93.28%

formance of the Classifier, equations (5) (9) in Section 2.5.2 were calculated the recall, specificity, precision, accuracy, and F1-score for 4 categories as Chewing, Swallow, Talk, and Other, shown as Table 6.

The accuracy rate for Swallow is 73.64% with the lowest value and the accuracy of Ture label for Chewing and predicted as Other was 21.62%. Which showed that this model is still in a difficult state to classify swallowing data perfectly. Several reason leads to the low accuracy of swallow data compared to other sound events. First reason is that the number of swallowing data in this datasets is lowest. In recorded data, approximately 15 or more chewing event corresponding to one swallow event. Therefore, the total number of swallow sound clip data is only 1129, 2.5% of the total sound clip data number (44595), 10% of the total chewing sound clip data number (11304). This may result in an imbalance in the data. To account for this, we incorporated the weighting of the LSTM model. However, we need to further improve the balance of the data.

4 DISCUSSION

4.1 Performance Comparison with Previous Research

Several classification method for detection of eating behavior have been developed in the last decade. Although the technology of dietary monitoring is evolving rapidly with the advancement of sensors, automatic monitoring of comprehensive dietary intake in real time is still one of the big challenge worth studying. There are two ways to classify eating behavior. One is to classify whether it is during the meal period or not, or to classify chewing, swallowing, and speech in real time as in this study.

There are already a large number of dietary period classifications in existence and with a sufficiently high level of accuracy. Gao et al. develop a practical solution for automatic detection of eating episode (Gao et al., 2016). They achieve the detection accuracy rate over 94% using LOSO (Leave One Sample Out) method based on deep learning. Bi et al. propose a wearable system for eating detection in free-living scenarios and achieve the accuracy rate over 90% but also belongs to the classification by dietary period (Bi et al., 2017). However, this kind of classification method cannot classify how many times one eating event has chewed during the meal period though the high accuracy. Zhang et al. used the same policy as in this study to achieve real-time classification of eating episodes (Zhang et al., 2020). They present the design, implementation, and evaluation of a necklace suit for detection and validation of chewing sequences and eating episodes in free living condition. They achieve the F1-score at 76.2% on per-second level and 81.6% at the per-episode level. Overall F1-score of 73.7% in detection the chewing sequences in a free-living condition. Kamachi et al. suggested an automatic segmentation method to detect eating activity using bone conduction sound the same as this study and porpoised a segmentation method based on the chewing model (Kamachi et al., 2021). The result comparison with all studies above and related study on chewing sound detection shown in Table 8. From the result comparison, our study performed a better F1-score in overall sound episode. Also in this study, the data was collected in a free-living condition too with talking and noises.

Table 7: Comparison between the developed YAMNet featured LSTM classifier and previous classifier.

Author	*Recorder*	*Events*	$F1Score_{comp}$
(*Bedriet al.*,2017)	*Mu*	*Ch,Dr,Ta,Sw,Wa*	80%
(*Gaoet al.*,2016)	*AR*	*Ea*	*Ac*_94%
(*Zhanget al.*,2020)	*IMU*	*Ea*	81.6%
(*Diouet al.*,2017)	*AR*	*Ch*	88.3%
(*Kamachiet al.*,2021)	*AR*	*Ch,Sw,Ta,O,Ea*	*Pre*_88.1%
Our method	*AR*	*Ch,Sw,Ta,O*	93.28%

Note: Where, $F1Score_{comp}$ means comprehensive F1-score; *Mu* means Multimodal sensor; *AR* means Audio record, *Ch* means Chewing event; *Dr* means Drinking event, *Ta* means Talking event; *Wa* means Walking event; *Ea* means Eating episode; *O* means Other(silent) data; *Sw* means Swallow event; *Ac* and *Pre* means overall accuracy and precision when F1-score is not provide, respectively.

4.2 Future Work for Classification

As indicated by the results in this study, the swallow data holds a low percentage of accuracy rate. The addition of sound episodes is becoming essential in the future to improve the classifier. It is also possible that the volume of swallowing sounds taken with bone conduction microphones varies greatly and that sound episodes of swallowing that are too small were not adequately captured in the labeling phase of this study. The future plan designed as following:

- Eliminate artificial error during labeling by employing throat microphones in addition to bone conduction microphones and video in datasets collection.
- Classification of eating behavior by simultaneous input of throat microphone and bone conduction microphone when classification need better swallowing sound estimation result.

5 CONCLUSIONS

In this study, we developed a classification system for human eating behavior in a natural environment using YAMNet transition learning and LSTM networks. The data for the classification of eating behavior consists of chewing, swallowing, speech, and other signals including noises. In particular, we found that the use of transition learning in YAMNet can enhance the features of speech signals and improve the accuracy rate of classification models for machine learning and Deep Learning. Using LSTM, we built a classifier using the embedding layer of YAMNet as the input feature. The F1-score and accuracy rate of the overall classified data reached 93.28% and 93.3%, respectively. By using the classification prediction of this research, we can canonically estimate the number of chewing and swallowing in real time, and expect to build a smarter eating environment by digitizing eating behavior and preventing fast eating.

ACKNOWLEDGEMENTS

This research was supported by Lotte Research Promotion Grant. Entire experimental protocols were approved by the ethics committee of Aoyama Gakuin University (H18-003).

REFERENCES

Altché, F. and de La Fortelle, A. (2017). An lstm network for highway trajectory prediction. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 353–359. IEEE.

Amft, O. (2010). A wearable earpad sensor for chewing monitoring. In *SENSORS, 2010 IEEE*, pages 222–227. IEEE.

Bae, S. H., Choi, I., and Kim, N. S. (2016). Acoustic scene classification using parallel combination of lstm and cnn. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, pages 11–15.

Bedri, A., Li, R., Haynes, M., Kosaraju, R. P., Grover, I., Prioleau, T., Beh, M. Y., Goel, M., Starner, T., and Abowd, G. (2017). Earbit: using wearable sensors to detect eating episodes in unconstrained environments. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(3):1–20.

Bi, S., Wang, T., Davenport, E., Peterson, R., Halter, R., Sorber, J., and Kotz, D. (2017). Toward a wearable sensor for eating detection. In *Proceedings of the 2017 Workshop on Wearable Systems and Applications*, pages 17–22.

Chun, K. S., Bhattacharya, S., and Thomaz, E. (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.

Clark, J. P., Adams, S. G., Dykstra, A. D., Moodie, S., and Jog, M. (2014). Loudness perception and speech intensity control in parkinson's disease. *Journal of communication disorders*, 51:1–12.

Diou, C., Papapanagiotou, V., and Delopoulos, A. (2017). Chewing detection from an in-ear microphone using convolutional neural networks. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1258–1261. IEEE.

Dong, X., Yin, B., Cong, Y., Du, Z., and Huang, X. (2020). Environment sound event classification with a two-stream convolutional neural network. *IEEE Access*, 8:125714–125721.

Elizalde, B. and Tompkins, D. (2021). Covid-19 detection using recorded coughs in the 2021 dicova challenge. *arXiv preprint arXiv:2105.10619*.

Gao, Y., Zhang, N., Wang, H., Ding, X., Ye, X., Chen, G., and Cao, Y. (2016). ihear food: eating detection using commodity bluetooth headsets. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 163–172. IEEE.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.

Hochreiter, S. and Schmidhuber, J. (1997). Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, pages 473–479.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, Q., Wang, W., and Zhang, Q. (2017). Your glasses know your diet: Dietary monitoring using electromyography sensors. *IEEE Internet of Things Journal*, 4(3):705–712.

Ittichaichareon, C., Suksri, S., and Yingthawornsuk, T. (2012). Speech recognition using mfcc. In *International conference on computer graphics, simulation and modeling*, pages 135–138.

Kamachi, H., Kondo, T., Hossain, T., Yokokubo, A., and Lopez, G. (2021). Automatic segmentation method of bone conduction sound for eating activity detailed detection. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pages 310–315.

Kamachi, H., Kondo, T., Yokokubo, A., and Lopez, G. (2020). Classification method of eating behavior by dietary sound collected in natural meal environment. In *Activity and Behavior Computing, ABC 2020. Smart Innovation, Systems and Technologies*, volume 204, pages 135–152. Springer, Singapore.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, J., Park, J., Kim, K. L., and Nam, J. (2017). Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *arXiv preprint arXiv:1703.01789*.

Navarro, J. M., Martínez-España, R., Bueno-Crespo, A., Martínez, R., and Cecilia, J. M. (2020). Sound levels forecasting in an acoustic sensor network using a deep neural network. *Sensors*, 20(3):903.

NCD Risk Factor Collaboration (2016). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19· 2 million participants. *The lancet*, 387(10026):1377–1396.

Nkurikiyeyezu, K., Kamachi, H., Kondo, T., Jain, A., Yokokubo, A., and Lopez, G. (2021). Classification of eating behaviors in unconstrained environments. *Biomedical Engineering Systems and Technologies, BIOSTEC 2020. Communications in Computer and Information Science*, 1400:592–609.

Ntalampiras, S., Kosmin, D., and Sanchez, J. (2021). Acoustic classification of individual cat vocalizations in evolving environments. In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pages 254–258. IEEE.

Päßler, S. and Fischer, W.-J. (2011). Acoustical method for objective food intake monitoring using a wearable sensor system. In *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 266–269. IEEE.

Prakash, J., Yang, Z., Wei, Y.-L., Hassanieh, H., and Choudhury, R. R. (2020). Earsense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–13.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

Selamat, N. A. and Ali, S. H. M. (2020). Automatic food intake monitoring based on chewing activity: A survey. *IEEE Access*, 8:48846–48869.

Semeniuta, S., Severyn, A., and Barth, E. (2016). Recurrent dropout without memory loss. *arXiv preprint arXiv:1603.05118*.

Shen, Y., Salley, J., Muth, E., and Hoover, A. (2016). Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables. *IEEE journal of biomedical and health informatics*, 21(3):599–606.

Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., Quitry, F. d. C., Tagliasacchi, M., Shavitt, I., Emanuel, D., and Haviv, Y. (2020). Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764*.

Shuzo, M., Komori, S., Takashima, T., Lopez, G., Tatsuta, S., Yanagimoto, S., Warisawa, S., Delaunay, J.-J., and Yamada, I. (2010). Wearable eating habit sensing system using internal body sound. *Journal of Advance Mechanical Design, Systems, and Manufacturing*, 4(1):158–166.

Tensorflow (2020). Sound classification with yamnet. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet/, (Accessed:20 December 2021).

Totakura, V., Janmanchi, M. K., Rajesh, D., and Hussan, M. T. (2020). Prediction of animal vocal emotions using convolutional neural network. *International Journal of Scientific & Technology Research*, 9(2):6007–6011.

Vasudevan, H., Michalas, A., Shekokar, N., and Narvekar, M. (2020). *Advanced Computing Technologies and Applications: Proceedings of 2nd International Conference on Advanced Computing Technologies and Applications—ICACTA 2020.* Springer Nature.

Xu, Y., Kong, Q., Wang, W., and Plumbley, M. D. (2018). Large-scale weakly supervised audio classification using gated convolutional neural network. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 121–125. IEEE.

Yamaji, T., Mikami, S., Kobatake, H., Kobayashi, K., Tanaka, H., and Tanaka, K. (2018). Does eating fast cause obesity and metabolic syndrome? *Journal of the American College of Cardiology*, 71(11S):A1846–A1846.

Yang, X., Doulah, A., Farooq, M., Parton, J., McCrory, M. A., Higgins, J. A., and Sazonov, E. (2019). Statistical models for meal-level estimation of mass and energy intake using features derived from video observation and a chewing sensor. *Scientific reports*, 9(1):1–10.

Zeng, Y., Mao, H., Peng, D., and Yi, Z. (2019). Spectrogram based multi-task audio classification. *Multimedia Tools and Applications*, 78(3):3705–3722.

Zhang, H., Lopez, G., Shuzo, M., Delaunay, J.-J., and Yamada, I. (2011). Analysis of eating habits using sound information from a bone-conduction sensor. In *Proc. of the IADIS e-Health Conference (EH 2011)*.

Zhang, J. and Man, K.-F. (1998). Time series prediction using rnn in multi-dimension embedding phase space. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, volume 2, pages 1868–1873. IEEE.

Zhang, S., Zhao, Y., Nguyen, D. T., Xu, R., Sen, S., Hester, J., and Alshurafa, N. (2020). Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–26.

Zheng, X. and Yan, J. (2019). Acoustic scene classification combining log-mel cnn model and end-to-end model. *DCASE2019 Challenge*.