A Data Augmentation Technique for Automatic Detection of Chewing Side and Swallowing

Akihiro Nakamura*, Takato Saito†, Daizo Ikeda†, Ken Ohta†, Hiroshi Mineno*, Masafumi Nishimura*
*Shizuoka University, Shizuoka, Japan
E-mail: nakamura.akihiro.16@shizuoka.ac.jp, mineno@inf.shizuoka.ac.jp, nisimura@inf.shizuoka.ac.jp
†Service Innovation Department, NTT DOCOMO. Inc, Tokyo, Japan
E-mail: takato.saitou.bu@nttdocomo.com, ikedad@nttdocomo.com, ootaken@nttdocomo.com

***Abstract*—Poor quality of eating behavior is known to have adverse effects on health. With a view to promoting health, this study examines a monitoring system for eating behavior that uses a convenient microphone. We previously performed automatic detection of masticatory balance and swallowing using two-channel microphone recordings and the Hybrid CTC/Attention Model to detect the quality of eating behavior. In this paper, we propose an N-gram based data augmentation technique using a large amount of weakly labeled data to improve the accuracy of automatic detection. The application of this method to the Hybrid CTC/Attention Model resulted in improved detection performance. Moreover, the performance of open foods not included in the training data was shown to be similar to that of closed foods.**

I. Introduction

Good eating behavior is vital to human health, and chewing and swallowing are particularly important. People with low chewing frequency per swallow tend to eat fast, which can lead to obesity and lifestyle-related diseases. Moreover, partial chewing causes tooth loss. Therefore, to maintain health, it is important to monitor eating behavior daily.

As techniques for monitoring eating behavior, the Inertial Measurement Unit (IMU) and image are widely used. Studies using IMU have included the detection of food intake [1], estimation of swallowing frequency [2], and estimation of chewing frequency [3]. Then, studies using image have included the detection of dietary behavior [4] and detection of chewing [5]. However, the use of IMU can pose a burden on the body if the instrument is worn for a long time, and this is a barrier to daily monitoring. On the other hand, the use of the image presents issues of privacy.

Therefore, it is worth exploring new techniques that use a convenient microphone to monitor eating behavior. For example, Olanjo et al. [6] used a throat microphone to estimate the number of swallows. Yin et al. [7] used a throat microphone to estimate food and drink intake and to classify solids and liquids. Moreover, methods using RNN such as LSTM have been employed to automatically detect feeding behavior. Ando et al. [8] used the Gaussian Mixture Model (GMM) to identify chewing, swallowing, and speaking in frames, and LSTM combined them to estimate eating behavior. However, an accurate label of the frame unit (strong label) was necessary for the training data. It was difficult to collect large amounts of strong label data using LSTM, and sufficient performance could not be obtained with a small amount of data. Billah et al. [9] showed that using Long Short-term Memory-Connectionist Temporal Classification (LSTM-CTC) could significantly improve the performance of chewing and swallowing detection by utilizing a large amount of training data with only weakly labeled data without accurate time information that could be collected relatively easily.

On the other hand, these studies were limited to the simple detection of the number of times and the estimation of eating behavior; they did not address the detection of the quality of eating behavior. Therefore, we previously attempted to automatically detect chewing position (front/left/right) and swallowing using two-channel recorded sounds and the Hybrid CTC/Attention Model with weakly labeled training data to evaluate the balance of chewing and swallowing [10]. However, in order to train the context by Attention, a large amount of training data is needed, and the financial cost of recording a large amount of food sounds is high.

In this study, we propose an N-gram based data augmentation method using weakly labeled data. The method generates a large amount of weakly labeled eating sounds. We apply this method to the Hybrid CTC/Attention Model to test whether it improves detection performance. Further, we evaluate whether the performance of open foods that are not included in the training data is similar to that of closed foods.

II. Hybrid CTC/Attention Model

There are two kinds of end-to-end methods that can learn using weakly labeled data without accurate time information: Connectionist Temporal Classification (CTC) [11] and the Encoder-Decoder model using Attention [12]. In addition, there is the Hybrid CTC/Attention Model using the CTC and the Attention [13].

A. Connectionist Temporal Classification (CTC)

CTC is a loss function developed for constructing an end-to-end speech recognition system that directly outputs words and sentences from acoustic features in the speech recognition field. In CTC, a blank label called “blank” is introduced between

symbols, each symbol is allowed to be output continuously, and the length of input and output is made to coincide to obtain consistency, so that training by weakly labels is enabled.

Let $\boldsymbol{x} = \{x^1, x^2, \cdots, x^T\}$ be the input data sequence, $\boldsymbol{y} = \{y^1, y^2, \cdots, y^T\}$ be the output of the network, and $\boldsymbol{l} = \{l^1, l^2, \cdots, l^{T'}\}$ be the correct label sequence. Generally, the length T' of $\boldsymbol{l}$ is shorter than the length T of the input data sequence. A redundant label sequence $\boldsymbol{\pi} = \{\pi^1, \pi^2, \cdots, \pi^T\}$ corresponding to $\boldsymbol{l}$ and a function B for mapping $\boldsymbol{\pi}$ to $\boldsymbol{l}$ are defined. Function B removes the same sequence of labels as "blank" and returns the final phoneme sequence. Then, $p(\boldsymbol{\pi}|\boldsymbol{x})$ is expressed by (1).

$$p(\boldsymbol{\pi}|\boldsymbol{x}) = \prod_{t=1}^{T} y_{\pi_t}^t \tag{1}$$

The probability $p(\boldsymbol{l}|\boldsymbol{x})$ that the input/output sequence is the label sequence $\boldsymbol{l}$ is expressed by (2) using $p(\boldsymbol{\pi}|\boldsymbol{x})$.

$$p(\boldsymbol{l}|\boldsymbol{x}) = \sum_{\boldsymbol{\pi} \in B^{-1}(\boldsymbol{l})} p(\boldsymbol{\pi}|\boldsymbol{x}) \tag{2}$$

Using the maximum likelihood estimation, the parameter that maximizes $p(\boldsymbol{l}|\boldsymbol{x})$ is obtained, and the training model is created.

B. Encoder-Decoder Model Using Attention

Attention is a framework introduced in the Encoder-Decoder model. The Encoder takes input and generates a fixed length vector $\boldsymbol{h} = \{h^1, h^2, \cdots, h^{T'}\}$ in Long Short-Term Memory (LSTM). In Decoder, the output $\boldsymbol{y} = \{\boldsymbol{y^1}, \boldsymbol{y^2}, \cdots, \boldsymbol{y^{T''}}\}$ is generated by LSTM from the vector generated by Encoder. In Decoder, the t-th hidden state $\boldsymbol{q^t}$ is calculated as shown in (3) and (4),

$$\boldsymbol{q^t} = \mathrm{LSTM}(\boldsymbol{r^t}, \boldsymbol{q^{t-1}}, \boldsymbol{y^{t-1}}) \tag{3}$$

$$\boldsymbol{r^t} = \sum_{i=1}^{T} \alpha_{t,i} \boldsymbol{h_i} \tag{4}$$

where $\alpha_{t,i}$ is the weight of the Attention at time t, which means the rate of utilization of the hidden layer state $\boldsymbol{h_i}$ at each time in the Encoder. This is unlike with CTC, where it was assumed that each event occurred independently. Instead, by using Attention, it is possible to reflect the history of past output and to train a series of contexts.

C. Hybrid CTC/Attention Model

Watanabe et al. [13] reported that the Hybrid CTC/Attention Model achieved drastic improvement in the accuracy of speech recognition. In an experiment using the Corpus of Spontaneous Japanese (CSJ), the character error rate (CER) improved 9.4% with CTC only, 11.4% with Attention only, and 8.4% with the hybrid model.

In the Hybrid CTC/Attention Model, the final output is the sum of the vectors output by CTC and Attention, and there is an advantage that features of both CTC and Attention are utilized. The training loss function L is defined by the weighted linear sum of the loss functions of CTC and Attention, as expressed by (5).

$$\boldsymbol{L} = \alpha \boldsymbol{L_{CTC}} + (1 - \alpha) \boldsymbol{L_{Attention}} \tag{5}$$

We previously applied this model to detect chewing and swallowing [10]. In addition to the detection of each eating behavior in CTC, the accuracy improvement of event detection can be expected using Attention which can reflect the history of complicated eating behavior. Fig. 1 shows an example of an event sequence when a cracker is eaten. Initial prechewing occurs, followed by repeated bilateral chewing before swallowing. The sequence including chewing position and swallowing is found to be a complex sequence of events.

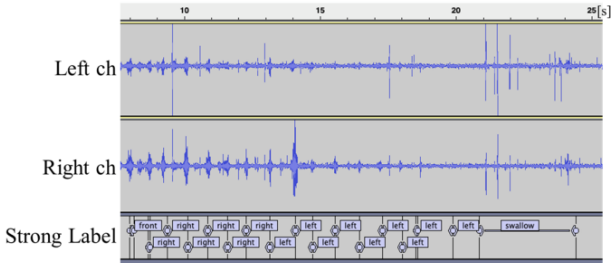

Fig. 1. Example of a recorded eating sound (two-channel) and event sequence (strong label) given manually for one cracker

III. Automatic Detection System

From the collected eating sounds, each event was classified into one of five classes: left chewing, right chewing, front chewing, swallowing, or others.

First, the eating sound was recorded using a two-channel condenser microphone placed under the ear (16 bit, 22 KHz sampling). Fig. 2 shows the installation and microphone unit. The microphone was created independently with a 3D printer. In addition, we assigned weakly labels for each event to create the training model. The weakly labels did not have accurate time information. Labeling was done by an online application created to reduce the cost of labeling and the log generated by the subject pressing a key during each chewing or swallowing event.

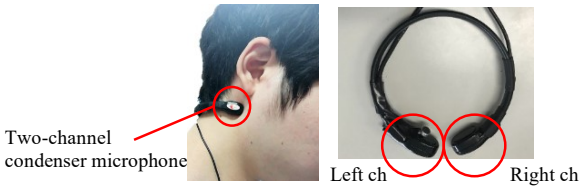

Fig. 2. Installation of the microphone and the microphone unit

Next, the recorded eating sound was converted into a voice feature amount. It was extracted in the window width of 80 ms and the frame shift of 40 ms. In addition to the 39-dimensional MFCC extracted by adding the left and right signals for signal enhancement, the feature value was

obtained by adding the seven-dimensional cross-correlation value to improve the detection performance for chewing position [8]. The MFCC consisted of 12 units with 1 dimensional Root Mean Square (RMS), 13-dimensional Δ, and 13-dimensional ΔΔ.

This feature was input to the Hybrid CTC/Attention Model, an automatic detector, to estimate chewing position and swallowing. When the SoftMax function was applied to the network output, double threshold [14] was applied as smoothing to prevent the chewing alternation within a short period of time. The threshold values φ_{low} and φ_{high} were decided, and the output probability over φ_{high} was marked. Then, frames larger than φ_{low} adjacent to the marked prediction were retrieved, and these were considered valid events. When a change in chewing occurred in a short period of time, this treatment was effective in correcting the change so that chewing occurred on one side.

IV. Data Augmentation

This section describes the data augmentation method using recorded eating sounds and weakly labeled data.

First, as shown in Fig. 3, an automatic detection system was constructed using weakly labeled eating sounds for training and CTC. Next, these eating sounds were input to the system, and alignment estimation by CTC was performed for each chewing (front/left/right) and swallowing event. The speech waveform corresponding to the event detected by the estimation was collected, and a database was generated. Each event waveform took the interval from the start of the corresponding alignment to the start of the next alignment.

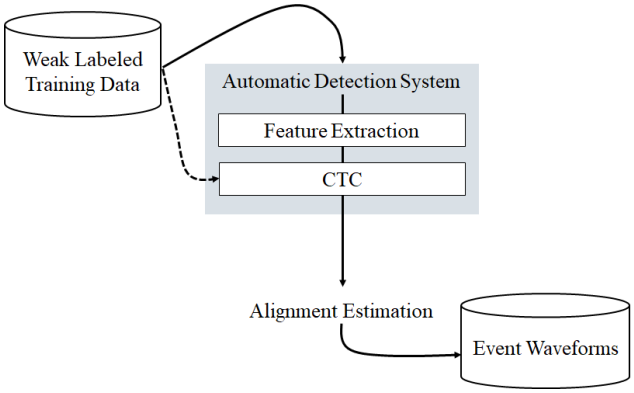

Fig. 3. Division of event waveform using weakly labeled data

Next, as shown in Fig. 4, the training data were augmented using the event waveforms collected by alignment estimation and the 5-gram model. The 5-gram model is based on the chewing side (front/left/right) label, which is a weakly label included in the training data. The next chewing type (front/left/right) was determined repeatedly by generating random numbers with probability based on 5-gram, and an event sequence was generated. Concerning swallowing, a normal distribution was assumed for the distribution of chewing frequency. In the training data, the average number of chews from the start of eating to swallowing μ and the standard deviation σ were obtained. Then, a normal distribution random number was obtained, and this was used as the number of chewing events until swallowing occurred.

Finally, eating sounds were generated by selecting and concatenating waveforms corresponding to respective events of the generated event sequence from candidates. For the waveform to be used in this case, a waveform having a similar position from the head should be selected as much as possible. Let n be the number of candidate event waveforms to be concatenated and t_i be the position from the head of the i ($1 \leq i \leq n$)-th event waveform. The position from the beginning of the eating sound currently being generated is defined as c. The i-th event waveform was randomly selected at a rate of $\frac{1}{|t_i - c|+1}$ out of all the candidates. By selecting the waveform in which the time was close, the difference of the chewing sound by the change of the shape of the food material by the chewing was reflected. By selecting the one close to the current time, it was possible to maintain the feature that the chewing sound in the first half before the food was crushed was large, and the chewing sound in the second half was small.

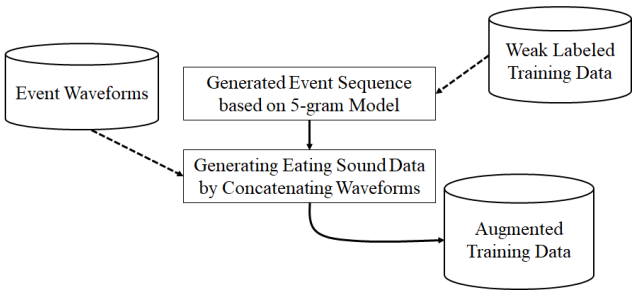

Fig. 4. Data augmentation by concatenating waveforms

However, since the eating sounds differed by speaker and food, the data reinforcement described earlier was carried out for each speaker and each food type.

V. Evaluation

A. Dataset

As data for the training, the authors collected the food sounds of chewing gum, crackers (Ritz), and cabbage (shredded) for 26 men and 4 women in their twenties. In one recording, chewing gum was chewed for 3 minutes, and one cracker and 7 g cabbage were eaten. At the same time, weakly labels were applied based on the subjects' self-reports, and data were collected for a total of 29,835 chewing and 3,308 swallowing events. For the evaluation of detection performance, strong labels were manually attached to all data separately, but this information was not used for the training.

For the evaluation data for the open food, the authors recorded the eating sound of apples and pizza for 4 men and 1 woman in their twenties who had not provided the data for the training. In each recording, 12 g of apple and 4 cm square of pizza were eaten. Strong labels were applied 812 times for chewing and 40 times for swallowing.

B. Experimental Conditions

The Encoder of the Hybrid CTC/Attention Model uses unidirectional LSTM with two layers of 200 dimensions, and the Decoder uses unidirectional LSTM with one layer of 100 dimensions. The α which shows the ratio of the loss function of CTC and Attention was set to 0.7. The threshold values were set to $\varphi_{low} = 0.2$ and $\varphi_{high} = 0.75$.

Cross-verification was carried out by dividing the data for training of 30 persons into 6 sets by speaker, and the average of each result was obtained and evaluated. The model was constructed with the eating sounds from about 260,000 chewing and swallowing events with weakly labels, and eating sounds for all speakers and all foods were generated by data augmentation. The data augmentation was carried out for each speaker and each food according to the food, using all the eating sounds and weakly labels. The evaluation used the eating sounds and strong labels, including about 6,000 chewing and swallowing events that were not included in the training model.

The model used was the Hybrid CTC/Attention Model of the proposed method. We compared this model with CTC alone, where α = 1.0. The model with CTC alone does not consider context. For the data augmentation, we used only the eating sounds, and we also used 1, 3, 5, 10, and 15 times as the eating sounds. We compared the six types of data augmentation. Also, the training model of N-gram was be trained to generate a sequence of events during data augmentation compared to the 2-gram, 5-gram, 7-gram, and random.

Finally, to verify its practicality, we evaluated the detection performance using the evaluation data for apples and pizza. For each of the partitioned models, we evaluated the open foods and obtained the mean.

C. Evaluation Metric

The mean absolute percentage error (MAPE) estimated in the frame unit was used as an evaluation metric of the detection performance. MAPE was calculated by (6), where A_k is the number of correct answers, F_k is the number of estimates, and N is the number of data for evaluation.

$$\mathrm{MAPE} = \frac{100}{N}\sum_{k}^{N}\left|\frac{A_k - F_k}{A_k}\right| \quad (6)$$

We also evaluated Recall, Precision, and the F1 score of each event. It was considered that correct detection was performed when the overlap between the correct label and estimated label was detected.

D. Results

The detection performance of the whole event using closed foods is shown in Table 1. The model using Hybrid CTC/Attention which can consider the context showed much higher performance than the model using CTC alone. Moreover, by the data augmentation, there was the accuracy improvement in each model, and especially, the performance of Hybrid CTC/Attention Model, in which the training of the context advanced, was improved. Next, detection performance for each event using closed foods is shown in Table 2. Detection performance improved for all events. In particular, front chewing and swallowing, which had less data, were improved by data augmentation, and it is considered that training of the context of Attention was advanced by training with a large amount of data created by data augmentation. In addition, the detection performance of front chewing was lower than that of other events, but the accuracy could be improved by using other microphones in combination.

The relationship between the amount of data augmentation and the overall detection performance is shown in Fig. 5. In the case of data augmentation, the overall detection performance was improved in comparison to the case without data argumentation. Further, there was a plateau of the detection performance from 5 times to 10 times. Table 3 shows the results for when the N-gram used for data augmentation was changed. When 5-gram and 7-gram were used, the detection performance was improved and the effect of data augmentation was observed. On the other hand, when an event sequence was randomly generated or when 2-gram was used, the performance was lower than when data augmentation was not performed. The event sequences generated by these models did not reflect complex dietary histories and may have adversely affected the training of attention contexts.

Fig. 6 shows the relationship between the detection performance and chewing time for each food. For the verification of practicability, the evaluation was carried out with apple and pizza. The chewing time was normalized by dividing the time to swallow into five equal parts. In the early stage of chewing, before the food was ground, the chewing position could be distinguished with high accuracy. However, in the latter half of chewing, the detection performance for foods other than chewing gum was lower than in the first half. The same tendency was observed for the model generated after data augmentation as for the model without data augmentation [10]. The results seemed to be affected by the fact that the food was crushed and spread in the mouth, then chewed in the whole. Since the correct label gives the chewing position by the subject's self-report, the information of correct answers itself becomes ambiguous.

Table 4 shows the detection performance of each food. In the three-classes detection, the detection of the chewing position (front/left/right) was evaluated as simple chewing. Although there are some differences in the detection performance among foods, the detection of all foods including open foods is stable to some extent. The detection of mastication is more accurate

than the detection of the chewing position, and there are cases in which the detection of the chewing position is incorrect, although the detection of the chewing itself is possible.

TABLE 1. OVERALL DETECTION PERFORMANCE : FIVE-CLASS DETECTION (LEFT CHEWING, RIGHT CHEWING, FRONT CHEWING, SWALLOWING, OTHER)

Model	MAPE (%)
CTC Without Data Augmentation	31.3
CTC 10 × Data Augmentation, 5-gram	30.8
Hybrid CTC/Attention Without Data Augmentation	19.1
Hybrid CTC/Attention 10 × Data Augmentation, 5-gram	**18.6**

TABLE 2. DETECTION PERFORMANCE BY EVENT: FIVE-CLASS DETECTION (LEFT CHEWING, RIGHT CHEWING, FRONT CHEWING, SWALLOWING, OTHER)

Event	Hybrid CTC/Attention Without Data Augmentation			Hybrid CTC/Attention 10 × Data Augmentation 5-gram		
	Recall	Precision	F1 score	Recall	Precision	F1 score
Left Chewing	0.81	0.87	0.84	0.84	0.87	0.85
Right Chewing	0.82	0.86	0.84	0.82	0.88	0.85
Front Chewing	0.47	0.70	0.58	0.56	0.80	0.68
Swallowing	0.80	0.63	0.77	0.86	0.74	0.80

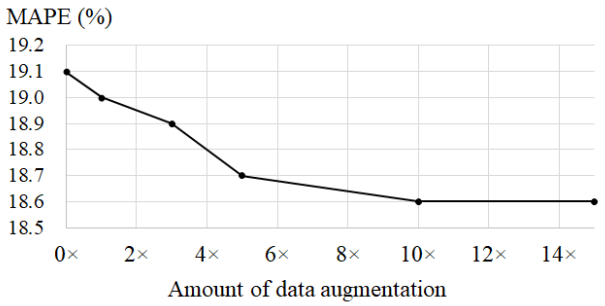

Fig.5. Relationship between the amount of data augmentation and overall detection performance (Hybrid CTC/Attention): five-class detection (left chewing, right chewing, front chewing, swallowing, other)

TABLE 3. N-GRAM AND OVERALL DETECTION PERFORMANCE (HYBRID CTC/ATTENTION): FIVE-CLASS DETECTION (LEFT CHEWING, RIGHT CHEWING, FRONT CHEWING, SWALLOWING, OTHER)

Model	MAPE (%)
Random	21.6
2-gram	19.8
5-gram	**18.6**
7-gram	**18.6**

TABLE 4 . DETECTION PERFORMANCE FOR CHEWING SIDE BY FOOD TYPE (HYBRID CTC/ATTENTION / 10 × DATA AUGMENTATION)

Food Type	THREE-CLASS DETECTION*			FIVE-CLASS DETECTION**		
	Recall	Precision	F1 score	Recall	Precision	F1 score
Gum	0.92	0.93	0.92	0.83	0.88	0.85
Cracker	0.91	0.95	0.93	0.81	0.89	0.85
Cabbage	0.93	0.95	0.94	0.82	0.88	0.84
Apple	0.89	0.95	0.92	0.82	0.89	0.86
Pizza	0.86	0.90	0.88	0.79	0.85	0.82

*** THREE-CLASS DETECTION: chewing, swallowing, other**
**** FIVE-CLASS DETECTION: left chewing, right chewing, front chewing, swallowing, other**

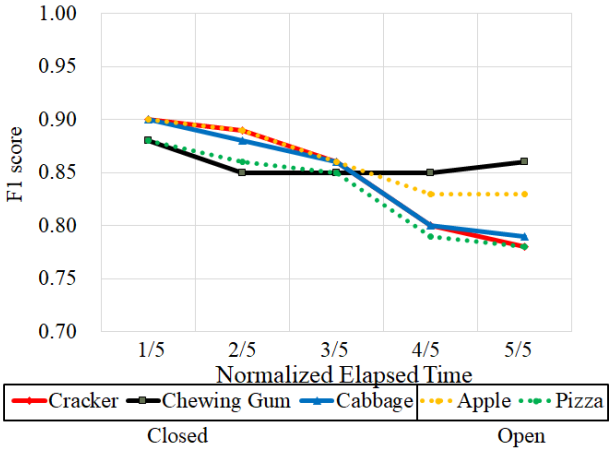

Fig.6. Relationship between normalized elapsed time and F1 score: five-class detection (left chewing, right chewing, front chewing, swallowing, other)

VI. CONCLUSION

The accuracy of the model was improved for closed and open foods by augmenting the training data with the proposed method. Specifically, compared with 19.1% of the MAPE of the Hybrid CTC/Attention model without data augmentation, 18.6% of the MAPE of the model that generated the event sequence based on the 5-gram model and augmented the data 10 times was obtained, showing the effect of data augmentation. The detection performance was improved especially for front chewing and swallowing, for which the amount of data was small. Moreover, as a result of evaluating open food, detection performance almost equivalent to closed food was obtained.

In the future, in order to improve the accuracy of the detection of chewing and swallowing, we will examine the possibility of using another model such as CRNN or Transformer, or using another microphone such as a sound collection microphone. In addition, we will work on the construction of a simple monitoring system for eating behavior, such as visualization of the number of times and the position of chewing.

ACKNOWLEDGEMENTS

This work was partially supported by JSPS KAKENHI Grant No. 18H03260.

REFERENCES

[1] Juan M. Fontana, Muhammad Farooq, Edward Sazonov, "Automatic Ingestion Monitor: A Novel Wearable Device for Monitoring of Ingestive Behavior," IEEE Trans Biomed Eng 2014, vol.61, no.6, pp.1772-1779, 2014.

[2] Dzung Tri Nguyen, Eli Cohen, Mohammad Pourhomayoun, Nabil Alshurafa, "SwallowNet: Recurrent Neural Network Detects and Characterizes Eating Patterns," 2017 IEEE International Conference on Pervasive Computing and Communications Workshops, pp.401-406, 2017.

[3] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, Gregory D. Abowd,"EarBit: Using Wearable Sensors to Detect Eating," IMWUT 2017, vol.1, no.3, 2017.

[4] Olubanjo Temiloluwa, Maysam Ghovanloo, "Real-time swallowing detection based on tracheal acoustics," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4384–4388, 2014.

[5] Xin Sun, Hongxun Yao, Wenyan Jia, Mingui Sun, "Eating activity detection from images acquired by a wearable camera," In Proceedings of the 4th International SenseCam & Pervasive Imaging Conference (SenseCam '13), pp.80-81, 2013.

[6] Fujimoto Atsuto, Ohkawauchi Takaaki, Yamato Junji, Ohya Jun, "An Image Processing Based Method for Chewing Detection Using Variable-intensity Template," Intelligent Robotics and Industrial Applications using Computer Vision 2018, pp.237-1-237-6, 2018.

[7] Bi Yin, Mingsong Lv, Chen Song, Wenyao Xu, Nan Guan, Wang Yi, "AutoDietary: A Wearable Acoustic Sensor System for Food Intake Recognition in Daily Life," IEEE Sens. J., vol.16, p.1, 2015.

[8] Jumpei Ando, Takato Saito, Satoshi Kawasaki, Masaji Katagiri, Daizo Ikeda, Hiroshi Mineno, Takashi Tsunakawa, Masafumi Nishida, Masafumi Nishimura, "Dietary and Conversational Behavior Monitoring by Using Sound Information," NCSP 2018, pp.675-678, 2018.

[9] Muhammad Mehedi Billah,Taiju Abe, Akihiro Nakamura, Takato Saito, Daizo Ikeda, Hiroshi Mineno, Masafumi Nishimura, "Estimation of Number of Chewing Strokes and Swallowing Events by Using LSTM-CTC and Throat Microphone," Proc. of GCCE 2019, pp.944-945, 2019.

[10] Akihiro Nakamura, Takato Saito, Daizo Ikeda, Ken Ohta, Hiroshi Mineno, Masafumi Nishimura, "Automatic Detection of the Chewing Side Using Two-channel Recordings under the Ear," Proc. of LifeTech 2020, pp.82-83, 2020.

[11] Graves Alex, Santiago Fernández, Faustino Gomez, Jürgen Schmidhuber, "Connectionist Temporal Classication:Labelling Unsegmented Sequence Data with Recurrent Neural Networks," Proc. Int. Conf. on Machine Learning, pp.369-376, 2006.

[12] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, "End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results," NIPS 2014 Workshop on Deep Learning, December 2014. 2014.

[13] Watanabe Shinji, Takaaki Hori, Suyoun Kim, John R. Hershey, Tomoki Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," IEEE Journal on Selected Topics in Signal Processing, vol. 11, no. 8, pp.1240-1253, 2017.

[14] Heinrich Dinkel, Kai Yu, "Duration Robust Weakly Supervised Sound Event Detection," ICASSP 2020, pp.311-315, 2020.