SSL-based Chewing and Swallowing Detection Using Multiple Skin-contact Microphones

Toshihiro Tsukagoshi*, Kazuhiro Koiwai*, Masafumi Nishida*, and Masafumi Nishimura*†
*Shizuoka University, Japan
†Aichi Sangyo University, Japan

***Abstract*—The recognition of eating behavior has a crucial role to play in the health management and monitoring of the elderly. Here, we propose an approach to the automatic detection of chewing and swallowing using skin-contact microphones. To improve the accuracy of the detection of chewing and swallowing, we explored the use of both self-supervised learning (SSL) and integrating information from multiple skin-contact microphones at different locations (2 channels below the ears and 2 channels on the throat). We evaluated our approach with training data collected from 27 subjects and test data from 5 subjects who did not contribute to the training data. We achieved a high-accuracy, with a character error rate of 2.88%, using an SSL-based model with a 4-channel summed signal. Furthermore, a frame-level F1-score evaluation showed a high degree of detection accuracy for both chewing and swallowing, achieving F1-scores of 0.954 for chewing and 0.97 for swallowing, with a 0.1 s allowance. These results demonstrate that SSL-based speech recognition models can be effectively applied to the detection of chewing and swallowing, even using a simple Early Fusion approach.**

I. Introduction

Eating behaviors such as chewing and swallowing provide important information for health management. Chewing activates extensive areas of the brain, producing improved cognitive function [1]. Conversely, insufficient chewing or the habit of eating quickly is a cause of lifestyle-related diseases and obesity [2]. In addition, age-related declines in swallowing function affect the nutrition and quality of life of the elderly, increasing the risk of aspiration pneumonia and choking. The importance of regular screening for swallowing function is recognized [3].

Non-invasive methods of recognizing eating behavior include the use of inertial measurement units (IMUs) and skin-contact microphones; further, Nguyen [4] proposed a method of estimating swallowing frequency using long short-term memory (LSTM) for data that are obtained from a neck-worn IMU. Eating behavior sounds have been used as a non-invasive and efficient method of recognizing eating behavior. Skin-contact microphones are robust against external noise, as they record from a microphone unit in direct contact with the skin, and they are able to record biological sounds such as chewing and swallowing more easily than close-talking microphones [5,6]. Nakamura et al. [7] proposed an approach combining recorded audio from multiple skin-contact microphones with an attention-based information integration method [8]. This model uses individual encoders and attention for each sound source, a common decoder, and, later, retrains using a hierarchical attention network, showing high-accuracy.

However, Nakamura et al.'s proposed method integrates information from multiple sources in the model, making it difficult to integrate with existing technologies for speech recognition. By contrast, if the effectiveness of an approach that integrates source data in advance, such as Early Fusion, can be verified, this would be easier to fuse with different types of speech processing technologies. For example, it would be possible to build an integrated system that simultaneously recognizes eating sounds and speech.

In recent years, models utilizing self-supervised learning (SSL) have been recognized as performing well in terms of speech recognition. Wav2vec2.0 [9] forms a type of SSL-based model reported to show high performance in various downstream tasks by pre-training using a large amount of unlabeled speech data and then fine-tuning it with a small amount of labeled data. It can be expected that the high-accuracy recognition of eating behavior is achievable through fine-tuning a pre-trained model that has been trained on a large amount of speech data, with labeled eating behavior data.

Conformer [10] is a model that integrates the global context-capturing ability of the Transformer architecture with the local feature-extraction capability of convolutional neural networks (CNN). Transformer excels in processing long-range dependencies, but has challenges in extracting local information. To overcome this, Conformer adopts a structure that effectively combines self-attention and CNN. Because Conformer also excels in processing complex time-series data as a speech recognition model, it may be suitable for the analysis of eating behavior sounds.

In this study, we propose effective methods for utilizing SSL and data obtained from multiple skin-contact microphones to improve detection accuracy for chewing and swallowing. As

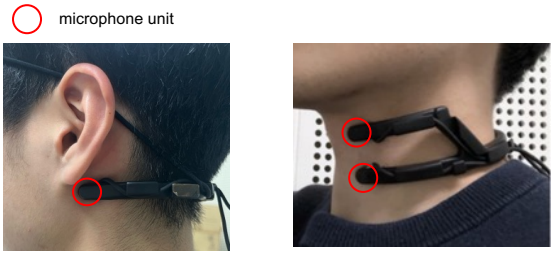

Fig. 1. Multiple skin-contact microphones

methods for utilizing multiple channels, we explored using input from individual channels as independent training data and also implemented a simple Early Fusion approach that sums multiple signals. Additionally, we consider data augmentation techniques. To evaluate the performance of this method, we conduct comparative experiments with the use of an open dataset respecting subjects, using Conformer as a comparison method. By means of these experiments, we seek to clarify the improvements in accuracy of the combination of SSL and multi-channel microphones in eating behavior detection and achieving a more accurate and robust detection system.

II. Approach

A. *Data processing*

To accurately record eating behavior, we placed skin-contact microphones below the ears on the left and right sides and upper and lower throat areas (Fig. 1). This arrangement was chosen to effectively capture the sounds of the main elements of eating behavior, namely, chewing and swallowing. As chewing involves the movement of both jaws, placing two-channel microphones on the left and right sides below the ears captures not only the differences in jaw movements but also the subtle changes in chewing sounds that are associated with position changes in food in the oral cavity. The microphones on the throat are primarily suitable for collecting swallowing sounds. Because, in swallowing, food flows from upper to lower, we placed two-channel microphones upper and lower throat areas. This arrangement allows us to capture a series of sounds going from the beginning to the end of swallowing, both temporally and spatially. In the recording environment, signals from the 4 channels are input into a single mixer, which produces parallel recorded audio data. In this study, we downsampled from the initial sampling rate of 22 to 16 kHz, which is suitable for inputting into speech recognition models. In addition, due to the characteristics of skin-contact microphones, low-frequency noise may occur due to blood flow in the neck area. To address this, we used a 100 Hz high-pass filter for all 4-channel signals to remove unwanted low-frequency components. We recorded eating behavior sounds using cabbage, crackers, gum, and water. For cabbage and crackers, we recorded both chewing and swallowing sounds.

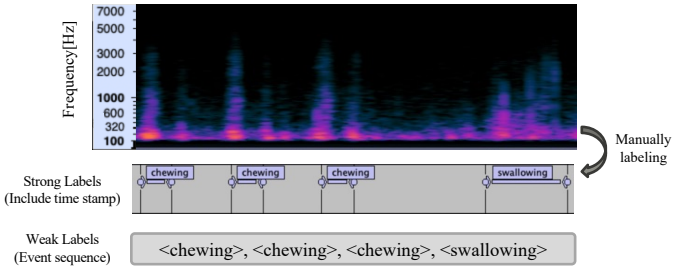

Fig. 2. Example of annotation for throat (upper) microphone

For gum, we recorded chewing sounds and saliva swallowing, and for water, we only recorded swallowing sounds.

We manually annotated strong labels that contained precise timing information for chewing and swallowing in eating behavior audio that was recorded with the 4-channel skin-contact microphones. Because speech recognition model training does not require exact timing information but only event sequences, we generated weak labels through omitting timing information from the strong labels. Fig. 2 presents an example of annotation for the throat (upper) microphone. We divided the audio data into segments of 10 seconds or less to maintain the time-series information of eating behavior and to ensure there was sufficient data for model training. When segmenting the audio, we referenced strong label data to avoid splitting swallowing or chewing events. This process ensures that each segment incorporates a complete sequence of eating behavior that is expected to improve model learning efficiency and recognition accuracy.

B. *Models*

For the creation of the SSL-based recognition model, we developed a model combining Wav2vec2.0, fully connected (FC) layers, and connectionist temporal classification (CTC). In this model, using Wav2vec2.0, we first extract 1024-dimensional features from the input audio data. The extracted features then pass through three FC layers (each layer's input and output dimensions are uniformly set to 1024), and finally connect to an FC layer with output nodes corresponding to the number of classification classes. The classification includes six classes, where <bos> and <eos> represent the beginning and end of a sentence, <unk> represents unknown events, <blank> represents blank spaces, and <chewing> and <swallowing> represent eating behaviors. For training, we use the CTC algorithm for loss calculation.

For comparison's sake, we prepared a model combining Conformer with CTC and KL divergence loss. In this model, we extract STFT features every 25 ms and apply an 80-dimensional mel-filterbank. The extracted features pass through two CNN layers before being input into the Conformer. The Conformer consists of 12 encoder blocks and 6 decoder blocks, and it finally generates a 6-dimensional output through an FC layer. The tokenizer used is the same as that in Wav2vec2.0. For training, we use a weighted combination of

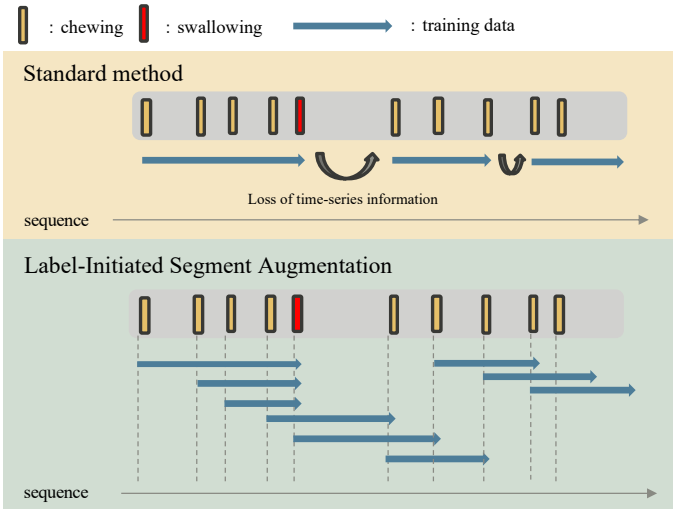

Fig. 3. Label-initiated segment augmentation

CTC and KL divergence loss functions. We utilized the speech recognition library SpeechBrain [11] for model training and evaluation.

C. *Data augmentation*

Unlike speech audio, eating behavior sounds feature a limited variety of labels, tending to have consecutive identical labels. Chewing labels make up the majority. The conventional method of dividing them into 10 s segments may result in a loss of time-series information for eating behavior at both ends of the segments. To address this, we propose a new data augmentation technique called Label-Initiated Segment Augmentation (LISA). LISA generates new training data through adopting each label as a starting point in addition to the existing dataset. As shown in Fig. 3, this method allows for a comprehensive incorporation of local time-series information included in the audio data to the training data. In addition, we introduced major audio augmentation techniques that are used in speech recognition model training. We applied Speed Perturbation [12], which brings about speed changes through resampling at a different sampling rate from the original waveform, Time Dropout [13], which zeros out random chunks of the original waveform, and Frequency Dropout [14], which zeros out signals in random frequency bands.

III. EXPERIMENTS

A. *Training data*

For training data, we used eating behavior audio drawn from 27 subjects (incorporating 17,539 instances of chewing and 1,329 instances of swallowing). To utilize 4-channel skin-contact microphone audio for learning, we considered two methods. The first is to sum audio signals. Taking the sum of these signals, we can integrate information from the microphones, and it is expected that the recognition model will produce more information. The second method is inputting each channel as independent training data for the model. Although information is not integrated at the input stage, it is expected that model accuracy can improve as data with different information used for learning. To verify the effects of the microphone placement, first, we used a single-channel method from the below-ear (left) microphone and single channel of the throat (upper) microphone as a baseline and then verified the combination using two channels of the below-ear microphone and two channels of the throat microphone. By means of these combinations of methods and placements, we comprehensively evaluated the effectiveness of each approach.

B. *Test data*

For the test data, we used eating behavior audio from five subjects who were not included in the training data (3,915 instances of chewing and 286 instances of swallowing). For the method using multi-channel summed audio for training data, we created and input-summed audio from the corresponding microphones for the test data.

C. *Training settings*

To create an SSL-based eating behavior recognition model, we used Wav2vec2-large-xlsr-53 as the pre-trained model. During the learning process, we updated not only the transformer blocks of Wav2vec2.0 but also the parameters for the feature extractor. For the Conformer model, we calculated the loss by combining CTC and KL divergence loss with a 3:7 weighting ratio. We used four Nvidia V100 SXM2 16 GB GPUs for model training. We used the correct weak labels from each dataset as targets to train each model. We allocated 10% of the training dataset for validation and continued training until the loss converged. We adopted the Adam optimizer in all models. For CTC decoding, we implemented beam search using a beam size of 10.

D. *Evaluation metrics*

The character error rate (CER) is widely used as an evaluation metric for speech recognition models. In this study, we also adopt CER for the evaluation criterion to measure the model's overall performance. In the context of the recognition of eating behavior, CER is suitable for evaluating classification accuracy in chewing and swallowing, in particular with regard to the accuracy of the chewing count. However, as chewing accounts for the majority of eating behavior, it is difficult to properly evaluate the recognition accuracy for swallowing using CER alone.

Thus, in this study, we used the F1-score for eating behavior detection with the use of strong labels as a second evaluation metric. This method adopts the frame-by-frame intermediate output of CTC. Usually, CTC compresses blanks following a frame-by-frame output, but, by analyzing the output before compression, a more detailed frame-by-frame evaluation is possible.

For evaluation, we used manually annotated strong labels as the ground truth for the test data, and we compared them with

Table 1. Comparison of recognition accuracy when using skin contact microphones (model: Wav2vec2.0)

	Training Data	*Test Data*	*LISA* *(Data Augmentation)*	*CER [%]*	*F1-score* *(Event: Chewing)*	*F1-score* *(Event: Swallowing)*
(1)	Below-ear: Left	Below-ear: Left	-	5.67	0.728	0.925
			○	3.27	0.754	0.933
(2)	Throat: Upper	Throat: Upper	-	13.7	0.661	0.948
			○	7.01	0.704	0.951
(3)	4ch-indipendent signal* (Below-ear, Throat)	Below-ear: Left	○	**2.69**	0.745	0.927
		Throat: Upper	○	3.76	0.720	0.935
(4)	4ch-summed signal (Below-ear, Throat)	4ch-summed signal (Below-ear, Throat)	○	2.88	**0.756**	**0.963**

* Each channel is input as independent training data for the model

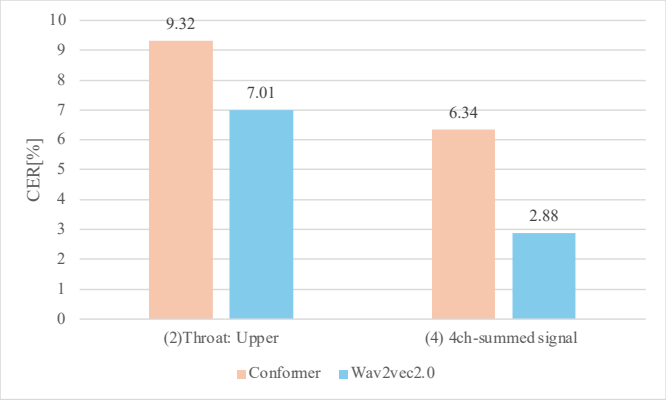

Fig. 4. Performance comparison between Conformer and Wav2vec2.0

the CTC frame output. In particular, we judged data to be as correct if the event frame for chewing or swallowing output by CTC was included in the range of the corresponding correct strong label. Furthermore, to explore the potential performance of the model, we set an allowance for the target labels and conducted evaluations at different tolerance widths.

E. Results

Table 1 presents the CER and F1-score for each training dataset with the Wav2vec2.0 model. The results of (1) show that LISA improved the accuracy for the Wav2vec2.0 model. This indicates that comprehensively capturing local time-series information improves the model's overall performance. A comparison of (1) and (2) shows that the detection accuracy of chewing was greater with the below-ear microphone and that of swallowing was higher using the throat microphone. The results of the F1-score showed that method (4) was the most accurate method for both chewing and swallowing. A comparison of the results of (3) and (4) showed that method (3) was the most accurate for CER values and method (4) was the most accurate for the F1-score. Because the CER value does not include time information, method (4) provides a more accurate output for time information. Further, the results showed that accuracy improved with the use of summed signals across 4 channels rather than their use as independent training data for the model. These results indicate that the left and right below-ear and upper and lower throat microphones provide distinct information, interactively providing improved model performance. Fig. 4 compares the performance of Wav2vec2.0 and Conformer. Conformer also showed performance improvement by using a 4-ch summed signal, but it did not exceed the accuracy of Wav2vec2.0. The results demonstrate that SSL-based speech recognition models can be effectively applied to the detection of chewing and swallowing.

Table 2. F1-score (Wav2vec2.0 + 4ch-summed signal)

Event	*Allowance[s]*			
	0	**0.01**	**0.05**	**0.1**
Chewing	0.756	0.798	0.905	0.954
Swallowing	0.963	0.963	0.970	0.970

Table 2 presents the results of the F1-score for the learning method, applying the 4ch-summed signal for the Wav2vec2.0 model, which provides the highest-accuracy. These results indicate that even without setting a time allowance, the recognition accuracy for swallowing behavior is very high. Further, setting only a 0.01 s allowance brings the F1-score for chewing to about 0.8, and with a 0.1 s allowance, the F1-score exceeds 0.95.

These results indicate that the proposed method can detect both chewing and swallowing with high temporal accuracy. Moreover, it was confirmed that it is possible to achieve high-accuracy detection of swallowing and chewing behaviors, even with a relatively simple Early Fusion approach using summed signals from multiple channels.

IV. CONCLUSION

In this study, we proposed effective methods for the use of SSL and data obtained from multiple skin-contact microphones to improve detection accuracy for chewing and swallowing. Comparative experiments using Wav2vec2.0 showed that an SSL-based speech recognition model is effective in detecting chewing and swallowing and that the relatively simple Early Fusion approach is also capable of highly accurate detection. We also showed that the newly proposed data augmentation technique LISA helps in the improvement of model performance.

Future works include developing further data augmentation techniques for swallowing-chewing sounds and the evaluation of the model in the context of more diverse eating patterns and

environments. Based on the results of this study, we also plan to extend this study to applied work, such as building an integrated system that simultaneously recognizes speech and eating behavior sounds.

ACKNOWLEDGMENTS

These results were obtained through a consignment project (JPNP 20006) by the New Energy and Industrial Technology Development Organization.

This research was partially supported by JSPS Grants-in-Aid for Scientific Research 18H03260 and 21K18305.

REFERENCES

[1] MOMOSE, T., et al. Effect of mastication on regional cerebral blood flow in humans examined by positron-emission tomography with 15O-labelled water and magnetic resonance imaging. *Archives of oral biology*, 1997, 42.1: 57-61.

[2] OTSUKA, Rei, et al. Eating fast leads to obesity: findings based on self-administered questionnaires among middle-aged Japanese men and women. *Journal of epidemiology*, 2006, 16.3: 117-124.

[3] HORIGUCHI, Satoshi, et al. Screening tests in evaluating swallowing function. *JMAJ*, 2011, 54.1:31-34.

[4] NGUYEN, Dzung Tri, et al. SwallowNet: Recurrent neural network detects and characterizes eating patterns. *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2017, 401-406.

[5] DUPONT, Stéphane, et al. Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise. *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, 2004.

[6] HERACLEOUS, Panikos, et al. Fusion of standard and alternative acoustic sensors for robust automatic speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, 4837-4840.

[7] NAKAMURA, Akihiro, et al. Automatic Detection of Chewing and Swallowing Using Attention-Based Fusion. *IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, 373-375.

[8] LI, Ruizhi, et al. Multi-stream end-to-end speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, 28: 646-655.

[9] BAEVSKI, Alexei, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020, 33: 12449-12460.

[10] GULATI, Anmol, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[11] RAVANELLI, Mirco, et al. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

[12] KO, Tom, et al. Audio augmentation for speech recognition. *Interspeech*, 2015, 3586.

[13] SRIVASTAVA, Nitish, et al. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014, 15.1: 1929-1958.

[14] ISLAM, Mobarakol; GLOCKER, Ben. Frequency dropout: Feature-level regularization via randomized filtering. *European Conference on Computer Vision. Cham: Springer Nature Switzerland*, 2022, 281-295.