J-010

会話を伴う日常食事環境における咀嚼音・嚥下音の自動認識
Automatic Recognition of Chewing and Swallowing Sounds in Conversational Daily Eating Environments

塚越 駿大† 西田 昌史† 西村 雅史†‡
Toshihiro Tsukagoshi Masafumi Nishida Masafumi Nishimura

1. はじめに

咀嚼・嚥下といった食行動は健康状態を示す重要な指標である。咀嚼回数の減少は肥満等の生活習慣病リスクを増大させ、嚥下機能の低下は誤嚥性肺炎等の疾患に直結する。食行動の自動認識システムが実現されれば、日常的な健康管理や疾病の早期発見・予防に向けたスクリーニングに貢献することが期待される。

これまでに非侵襲かつ自動的な食行動認識手法として、生体音を活用した咀嚼・嚥下認識技術が提案されており[1,2]、特に自己教師あり学習モデルを用いた場合に高い性能が報告されている。しかしながら、これらの研究は主に防音室等の管理環境での評価にとどまっており、実際の生活環境との乖離が課題となっている。日常的な食事環境では、複数人による会話音声、環境騒音、笑い声や嗳気等、実験室環境では発生しない多様な音が観測される。

本研究では、複数人の会話を含む日常的な食事環境を想定した実験系を構築し、自由な発話と食行動が混在する状況における既存モデルの性能を評価する。さらに、食行動音データに発話データを加えた学習データを用意し、これを使って既存のモデルをファインチューニングすることで、音声に対して頑健性を向上させる方法を提案する。

2. 提案手法

2.1 ベースライン: WavLM+GRU による食行動認識

WavLM+GRU モデルをベースラインとして使用した。このモデルは自己教師あり学習モデルである WavLM[3]を特徴抽出器とし、2 クラス{咀嚼,嚥下}の時系列分類に GRU を組み合わせた構成である。

2.2 音声データによるファインチューニング

ベースとなる WavLM+GRU モデルを食行動音のみで学習したのち、さらに音声データとして Commonvoice[4]日本語音声をこの学習データに加えた上で Fine tuning を行ったモデルを作成した(WavLM+GRU_CV)。出力は{咀嚼,嚥下,音声}の 3 クラス分類として学習を行う。発話音声を学習データに含めることで、食事中の会話に対する頑健性の向上が期待できる。

3. 評価実験

3.1 日常食事環境を想定した実験系の構築

複数人での会話を含む日常的な食事環境を模擬した実験系を構築した。試験環境を図 1 に示す。被験者は 3 名 1 組のグループで実験に参加し、相互に面識のある参加者同士で構成することで、自然な会話が発生しやすい環境を確保した。実験時間は 90 分間とし、初期 10 分間は摂食なし自由発話のみを行う環境で実施し、後半 80 分間では自由摂食および自由発話を行う環境で実施した。被験者には清涼飲料水およびスナック菓子を提供し、円卓を囲んで着席した状態でリラックスした摂食行動と会話を促した。これにより、実際の食事場面で頻繁に観測される、音声と食行動関連音の出現状況を再現した。

(a) 咽喉マイク (b) 耳下マイク

図 1 試験環境および皮膚接触型マイク

表 1. 評価用データセットの基本統計量

項目	平均値	標準偏差
咀嚼回数 (回/分)	16.4	0.9
嚥下回数 (回/分)	3.0	1.0
発話区間の割合 (%)	20.2	6.9

図 2 に示す皮膚接触型マイクを用いて、音声、咀嚼音、嚥下音等を含む生体音を収録した。同時に、被験者の正面にインターネットカメラを設置し、アノテーション作業時の同期映像として活用した。

3.3 学習用データセット

モデルの学習には防音室で収録された食行動時の生体音データを用いた。27 名の被験者がそれぞれ単独で収録に参加しており、発話などの音声は含まれず、食行動音のみで構成されている。キャベツ、クラッカー、ガム、水の摂取時に得られた咀嚼音 17,539 回、嚥下音 1,329 回が含まれる。

3.4 評価用データセット

モデルの評価には、2 種類のデータセットを使用した。

3.4.1 管理環境データ

3.3 節同様防音室で収録された生体音データのうち、学習に含まれない 5 名の被験者のデータを使用した。管理環境下におけるモデルの認識性能を確認するために用いた。

3.4.2 実環境データ

本研究で構築した日常食事環境の実験系において収録したデータのうち、摂食開始から 20 分間の食行動と会話が混在する区間を評価対象とした。被験者数は 3 名であり、いずれも学習データには含まれていない。データセットの統計情報を表 1 に示す。発話区間の割合は全体の約 20%であり、摂食行動については咀嚼回数が平均 16 回/分、嚥下回数が平均 3 回/分観測された。

† 静岡大学 Shizuoka University
‡ 愛知産業大学 Aichi Sangyo University

表 2. 管理環境データでの食行動認識精度

(a) 咀嚼イベント (IoU=0.01)

モデル	Precision	Recall	F1-score
WavLM+GRU	0.940	0.963	0.952
WavLM+GRU_CV	0.929	0.940	0.935

(b) 嚥下イベント (IoU=0.3)

モデル	Precision	Recall	F1-score
WavLM+GRU	0.927	0.937	0.932
WavLM+GRU_CV	0.881	0.915	0.898

表 3. 実環境データでの食行動認識精度

(a) 咀嚼イベント (IoU=0.01)

モデル	Precision	Recall	F1-score
WavLM+GRU	0.450	0.754	0.564
WavLM+GRU_CV	0.703	0.746	0.722

(b) 嚥下イベント (IoU=0.3)

モデル	Precision	Recall	F1-score
WavLM+GRU	0.564	0.805	0.657
WavLM+GRU_CV	0.799	0.797	0.797

3.5 評価指標

咀嚼および嚥下イベントの検出精度を F1 スコアにより評価した。イベント検出の評価には、予測区間と正解区間の重複度を示す IOU（Intersection over Union）しきい値を用いた。IOU しきい値は咀嚼 0.01（発生検出重視）、嚥下 0.3（区間精度重視）を設定した。また WavLM+GRU_CV モデルにおいて、入力音の違いによるモデルの動作特性を分析するため、誤認識が発生した事例を音響イベント別に分類した。具体的には発話、笑い声、息を吸う音といった各カテゴリにおける誤認識発生頻度を分析することで、モデルの限界と改善点を考察することを試みた。

4. 実験結果

表 2 および表 3 に、管理環境および実環境における食行動認識の各手法に対する F1 スコアを示す。ベースライン手法である WavLM+GRU は、管理環境下では咀嚼および嚥下のいずれにおいても高い F1 スコアを示したものの、実環境下では適合率の大幅な低下が見られた。具体的には、咀嚼検出における適合率は 0.45、嚥下検出では 0.56 にとどまり、音声区間における誤検出の多さが課題として浮き彫りとなった。一方、提案手法である WavLM+GRU_CV では、学習データに発話音声を加えることで、実環境下における誤検出の抑制に寄与した。咀嚼検出における適合率の向上により、F1 スコアは 0.56 から 0.72 へと大きく改善された。さらに、嚥下検出においても F1 スコアは 0.66 から 0.80 へと向上し、同様の効果が確認された。これらの結果から、発話が含まれる実環境においては、学習段階で発話音声を適切に反映することが、誤検出の抑制および認識精度の向上に有効であることが示唆された。

図 3 に、提案手法（WavLM+GRU_CV）による誤認識発生時の音響イベント分布を示す。誤検出が生じた区間の音響イベントを分析した結果、最も多く誤認識の原因となっていたのは「笑い声」であることが判明した。これは、学習データ中に笑い声が含まれていなかったことが主な要因と考えられる。さらに、発話音声によるファインチューニングを施した後でも、依然として発話に起因する誤認識は残存していた。相槌やフィラーといった短い発話は、咀嚼や嚥下と類似した音響特徴を有しており、これらが食行動として誤って検出されるケースが多く見受けられた。以上の結果から、今後より頑健な食行動認識モデルを構築するためには、従来の音声認識用コーパスに加えて、笑い声や自然な相槌・フィラー、を含む多様な音響イベントを学習データに網羅的に含めることの重要性が示唆される。

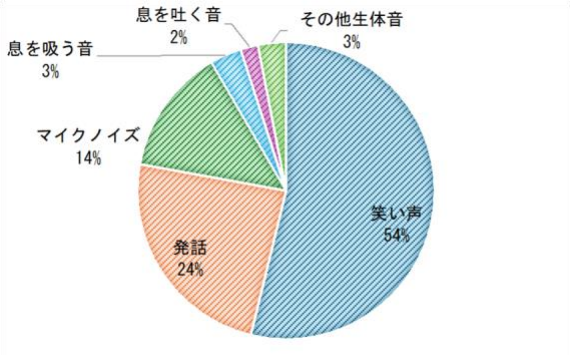

図 3. 誤認識が発生した音響イベントの内訳

5. おわりに

本研究では、会話を伴う日常食事環境において咀嚼・嚥下といった食行動をより高精度に認識するため、既存の認識モデルに音声データを加えてファインチューニングを行う手法を提案した。ベースライン手法が実環境で大きく精度を落とす一方で、音声を追加学習した手法では音声区間における誤検出を抑制し、咀嚼・嚥下のいずれにおいても F1 スコアが大きく改善された。この結果から、発話を含む多様な音響環境への適応において、発話データの事前学習が有効であることが示された。また、誤認識の主な要因として笑い声や相槌などが挙げられたことから、今後はこれらの音響イベントも学習に取り入れることで、さらに頑健な認識モデルを構築したい。

謝辞

本研究の一部は JSPS 科研費 18H03260, 21K18305 の助成を受けました。

参考文献

[1] Y. Khalifa, C. Donohue, J. L. Coyle, and E. Sejdic, “Autonomous swallow segment extraction using deep learning in neck-sensor vibratory signals from patients with dysphagia,” IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 2, pp. 956–967, 2023.
[2] Tsukagoshi, T.; Nishida, M.; Nishimura, M. Simultaneous Speech and Eating Behavior Recognition Using Data Augmentation and Two-Stage Fine-Tuning. Sensors 25, no. 5: 1544, 2025.
[3] Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process, 2022.
[4] Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common voice: A massively-multilingual speech corpus. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), Marseille, France, 2020.