5S-07

大規模事前学習モデルを用いた皮膚接触型マイク音声と食行動音の同時認識

小祝　和寛† 増田　光汰† 緒方　淳‡ 西田　昌史† 西村　雅史†

静岡大学†　　産業技術総合研究所‡

1. はじめに

高齢者の健康維持管理において「話す」「食べる」という二種類の行動が重要とされている．我々は皮膚接触型マイクである耳下マイクと咽喉マイクを用いて，音声や咀嚼音，嚥下音の収録を行なっており，これらの音情報を認識することで二種類の行動をモニタリングするシステムの検討を行っている．

これまでは，耳下マイクで収録した咀嚼音と嚥下音による咀嚼・嚥下の検出や，咽喉マイクで収録した音声による音声認識などを独立して検討してきた．一方で，同じ皮膚接触型マイクで収録されたこれらの音情報を一つのモデルで認識するという同時認識の研究には未着手であった．

このような観点から，今回我々は耳下マイクで収録された音声および食事音の同時認識の可能性を検証した．なお，弱ラベルによる学習に加え，データ数が少ない音情報を用いるため，大規模事前学習モデルを利用して認識モデルを構築した．

2. 提案手法

2.1. 耳下マイク

耳下装着の皮膚接触型マイクの装着例を図 1 に示す．この耳下マイクは主に咀嚼音と嚥下音を検出するために考案されたものであるが，咽喉部に装着した従来の咽喉マイクに近い品質の音声を収録できることがわかっている．このため，音声と咀嚼，嚥下などの生体音をこの一つのデバイスで全て同時に認識できる可能性があり，「話す」「食べる」といった機能を効率的にモニタリングする上で有効な手段になりうると考えた．

図 1 耳下 2ch マイクの装着例

Simultaneous Recognition of Skin-Contact Microphone Speech and Eating Behavior Sounds Using a Large-Scale Pre-Training Model

Kazuhiro Koiwai†, Kota Masuda†, Jun Ogata‡, Masafumi Nishida†, Masafumi Nishimura†

†Graduate School of Integrated Science and Technology, Shizuoka University

‡National Institute of Advanced Industrial Science and Technology

2.2. 事前学習モデルの活用方法

耳下マイクや咽喉マイクの音声は一般的な接話マイクよりも外部雑音に頑強であるが，高域が欠如するなど音響特性が異なるため，接話マイク音声で学習された音声認識システムでは高い精度を達成できない．また，耳下マイクや咽喉マイクの音声の大規模な学習用音声コーパスも存在しないため，専用のモデル構築が課題となっている．

増田ら[1]は，大規模事前学習モデルである wav2vec2.0 を用いて，少量のラベル付き咽喉マイク音声から咽喉マイク専用の音声認識モデルを構築し，接話マイク音声に近い認識精度を示した．

これより，同様の音響特性を持つ耳下マイク音声においても大規模事前学習モデルが適用できると考えた．本稿では，大規模事前学習モデルとして HuBERT を採用した．HuBERT は Hsu ら[2]によって提案された音声特徴表現フレームワークであり，wav2vec2.0 と同様に自己教師あり学習を用いた音声認識モデルである．この HuBERT は大規模事前学習モデルを用いた音声認識において wav2vec2.0 と同程度かそれを上回る精度を達成している．

HuBERT はモデルの内部で音データを特徴量に変換するため、入力として生の音声や食事音を用いる．ファインチューニング時には損失関数に CTC を用いており，時系列順にそれぞれのイベントを独立に検出できる．そのため，フレーム単位での正解ラベルである強ラベルは用いずに時系列情報の不要な正解弱ラベルを用いることで学習が可能であるため，咀嚼音と嚥下音においても弱ラベルを用いて学習を行う．

実験で用いる食事音のデータには，咀嚼の位置を考慮しない「咀嚼・嚥下」の 2 クラス分類でラベルが付与されている．今回の実験では咀嚼音を”1”，嚥下音を”2”とラベル付けすることで，同時に学習する音声データのラベルと区別してモデルの学習を行う．食事音のラベルの例を図 2 に示す．

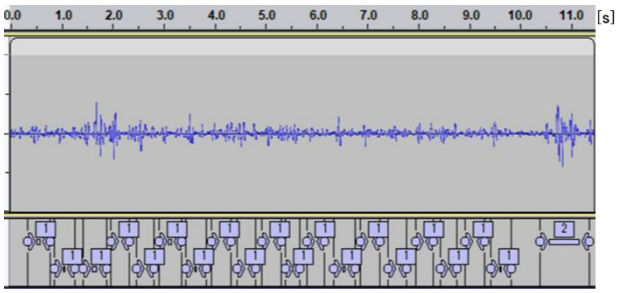

図 2 咀嚼(1)と嚥下(2)のラベルの例

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

3. 実験

3.1. 実験条件

データセットとして，耳下装着のマイクで収録された音声と食事音を学習データ，評価データで利用した．ファインチューニングで用いる学習データとして，男性 10 名から収録した音素バランス文読み上げ音声(約 5 時間)と，男性 15 名から収録したクラッカー(リッツ)・キャベツ(千切り)・ガムの食事音(約 140 分，咀嚼 11075 回，嚥下 431 回)を用いた．また，評価データとして，男性 5 名から収録した新聞記事読み上げ音声(約 25 分)と，男女 3 名から収録したクラッカー・キャベツ・ガムの食事音(約 25 分，咀嚼 1845 回，嚥下 113 回)を用いた．なお，評価データには学習データと同じ話者・被験者は含まれていない．また，どのデータにも弱ラベルが付与されており，評価データの食事音には強ラベルも付与されている．

ファインチューニングに用いる HuBERT の事前学習モデルとして，英語の大規模接話マイク音声コーパスの LibriLight60,000 時間で事前学習されたモデルである Large モデルを用いた．なお，このモデルは事前にファインチューニングされていない．

さらに，HuBERT の Large モデルに耳下マイク音声を学習データとしたものを音声の性能比較モデルとし，LSTM-CTC を用いた食事音専用モデルと食事音のみを学習データとした HuBERT の Large モデルを食事音の性能比較モデルとした．

評価尺度として，音声には CER を用いた．咀嚼音と嚥下音には強ラベルによる評価として MAPE(Mean Absolute Percentage Error)を用いた．MAPE では，データ数をN，正解回数をA_k，推定回数をF_kとして，以下の式で算出する．

$$MAPE = \frac{100}{N}\sum_{k}^{N}\left|\frac{A_k - F_k}{A_k}\right|$$

また，イベント別の検出性能として咀嚼と嚥下は F 値でも評価を行なっており，正解ラベルと推定ラベルの重なりが 1 点以上検出された場合に，そのイベントの正しい検出が行なわれたと判定している．

3.2. 実験結果

実験結果を表 1，表 2，表 3 に示す．

表 1 より，耳下マイク音声のみで学習されたモデルと CER が約 3%差という結果になったため，音声認識という観点では食事音との同時認識は大きな性能の低下にはつながらないことが分かった．

咀嚼・嚥下の検出において，表 2 より MAPE による全体の検出性能では LSTM-CTC と比べて，食事音のみで学習されたモデルも食事音と音声で同時に学習されたモデルのどちらも約 2 倍程度の値となった．一方で，表 3 より F 値によるイベント別の検出性能では，食事音と音声で同時に学習された HuBERT のモデルの咀嚼の検出率が他の 2 つのモデルに比べて約 0.2～0.25 ほど低くなった．嚥下に関しては食事音のみで学習された HuBERT モデルと同程度の結果となった．嚥下がどのモデルでも全体的に低くなっているのは，咀嚼音と音声に比べてデータ量が大きく不足していることが理由として考えられる．また，HuBERT で音声と同時に学習を行ったモデルでは，強い音情報である音声と同時に学習しているため，音声に比べて雑音に近い食事音の検出率が下がっていることが考えられる．

表 1 音声認識の結果

ベースモデル	学習データ	CER(%)
HuBERT	音声	13.8
HuBERT	音声+食事音	16.7

表 2 食事音における全体の検出性能

ベースモデル	学習データ	MAPE(%)
LSTM-CTC	食事音	29.6
HuBERT	食事音	53.6
HuBERT	音声+食事音	58.1

表 3 食事音におけるイベント別の検出性能

		F値(イベント別)	
ベースモデル	学習データ	咀嚼	嚥下
LSTM-CTC	食事音	0.52	0.33
HuBERT	食事音	0.47	0.19
HuBERT	音声+食事音	0.28	0.17

4. おわりに

今回の実験では，HuBERT を用いた耳下マイク収録の音声と食事音の同時認識の可能性を検証した．音声認識では大きな差はなかったが，食事音の検出に関してはまだまだ改善の余地があることが分かった．

今後は複数モデルの組み合わせによる同時認識や HuBERT による咀嚼・嚥下の検出の更なる工夫を実施していく．

謝辞

本研究の一部は JSPS 科研費 18H03260，21K18305 の助成を受けました．また，国立研究開発法人新エネルギー・産業技術総合開発機構(NEDO)の委託業務(JPNP20006)の結果得られたものです．

参考文献

[1] 増田光汰 ほか,:情報処理学会,2022
[2] Wei-Ning Hsu， et al．，“HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”， 2021

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.