E-016

大規模事前学習モデルとデータ拡張による音声・食行動音同時認識

Speech and Eating Sound Recognition Using Large-Scale Pretrained Model and Data Augmentation

小祝　和寛 [1)]　緒方　淳 [2)]　西田　昌史 [1)]　西村　雅史 [1)3)]
Kazuhiro Koiwai　Jun Ogata　Masafumi Nishida　Masafumi Nishimura

1. はじめに

我々は，高齢者の健康管理を目的として、耳下部に装着した皮膚接触型マイクで収録された音から「話す」「食べる」に関する行動情報を自動抽出するシステムの開発を行っている．

音声認識用のモデルとしては昨今，大規模事前学習モデルの有用性が広く知られており，皮膚接触型マイク音声についても，これらのモデルによって高い認識性能が得られることが知られている[1]．一方，食行動音のような特殊な音に対する適用可能性についてはまだ十分な検証は行われていない．なお，ここではシステム構成を簡素化するため，音声と食行動音を特に区別せず，同一の大規模事前学習モデルを用いて，両者を同時認識可能なモデルの構築を目指している．

我々は先に HuBERT を事前学習モデルとして音声と食行動音の同時認識を検討したが，十分な性能が得られなかった[2]ため，ここでは雑音を含んだ音声で事前学習された WavLM[3]に着目した．

また、Nakamura ら[4]が提案した N-gram による食行動音のデータ拡張手法にも着目した．この手法では，特に咀嚼音が食物粉砕の進行と共に変化する性質を考慮しながらデータ拡張を行っている．

本研究では，皮膚接触型マイク音声と食行動音及び拡張食行動音を用いて， WavLM へのファインチューニングを実施し，音声認識及び食行動音認識の観点から評価を行った．

2. N-gram を用いたデータ拡張

Nakamura らの手法では，まず弱ラベル食行動音データから LSTM-CTC によるアライメントを用いてイベント別食行動音データベースを作成している．イベント別食行動音データベースには，「前咀嚼・右咀嚼・左咀嚼・嚥下」に分類されてデータが格納されている．その後，弱ラベルで構築された 5-gram モデルを用いて，最初の咀嚼(噛み始め)から最後の嚥下までの一連の食行動音を表すイベント列を生成する．このイベント列に対応するように，イベント別食行動音データベースから音データを選択し連結して食行動音のデータ拡張を行っている．この時，食事の先頭からの位置が近い音データを選択することで，前半の咀嚼音は大きく，後半の咀嚼音は小さいという咀嚼音の特徴を反映した拡張食行動音データとなっている．

本研究では WavLM を認識モデルとして用いるため，WavLM の CNN Encoder を用いて食行動音の特徴量抽出を行った．また，「前咀嚼・右咀嚼・左咀嚼・嚥下」の 4 クラス分類ではなく，「前咀嚼・右咀嚼・左咀嚼」を「咀嚼」というラベルで統一し「咀嚼・嚥下」の 2 クラス分類とした．これは，チューニングデータを元の食行動音とする場合の弱ラベルや，評価データの食行動音の強ラベルにおいても同様である．

3. 認識モデルの構築

WavLM は事前学習時にマスクされた音声の予測と雑音の除去を学習するため，音声認識だけでなく多くの認識タスクにおいても優れた結果を残している事前学習モデルである．モデル構造は HuBERT をベースとしており，Transformer 部に対して Gated Relative Position Bias と呼ばれる機構を新たに導入している．今回用いる WavLM の Large モデルは，クリーンなデータである LibriLight60,000 時間に雑音を含むデータ 34,000 時間を加えた合計 94,000 時間の英語音声で事前学習されている．雑音を含んだ音声は，YouTube などから集められた英語音声データセットである GigaSpeech の 10,000 時間と，欧州議会の音声などで構成されているデータセットである VoxPopuli の 24,000 時間で構成されている．

ここでは，その Large モデルをファインチューニングすることで認識モデルの構築を行う．WavLM モデルより出力された文脈表現を線形層に通し，正解ラベルである弱ラベルとの CTC loss を最小化するように WavLM の Transformer 部ならびに線形層のパラメータを更新する．

4. 実験

4.1 実験条件

データセットとして，耳下マイクで収録された音声と食行動音をチューニングデータ，評価データで利用した．チューニングデータとして，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)・ガムの食行動音(約 2.3 時間，咀嚼

1) 静岡大学大学院総合科学技術研究科，Graduate School of Integrated Science and Technology, Shizuoka University
2) 産業技術総合研究所，National Institute of Advanced Industrial Science and Technology
3) 愛知産業大学，Aichi Sangyo University

11102 回，嚥下 435 回)を用いた．拡張手法を用いた認識モデルのチューニング時には，チューニング用の食行動音を 10 倍に拡張した 10 倍拡張食行動音(約 23 時間，咀嚼約 111020 回，嚥下約 4350 回)を用いた．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・ガムの食行動音(約 25 分，咀嚼 1845 回，嚥下 113 回)を用いた．なお，評価データにはチューニングデータと同じ話者・被験者は含まれていない．また，音声のチューニングデータ，評価データには平仮名による弱ラベルが付与されている．チューニング用食行動音及び 10 倍拡張食行動音には弱ラベルが付与されており，評価用食行動音には強ラベルが付与されている．

また，性能比較モデルとして，HuBERT の Large モデル及び LSTM-CTC モデルを用いる．HuBERT の Large モデルは LibriLight60,000 時間で事前学習されており，チューニングデータを用いてファインチューニングした．LSTM-CTC は 1 層 200 次元であり，データセットのチューニング用食行動音データを学習データに用いて，39 次元の MFCC でフレーム幅 80ms，シフト幅 40ms で特徴量を抽出しモデルの入力とした．

評価尺度として，音声認識では弱ラベルを用いた CER 評価を行った．食行動音認識では強ラベルを用いた F 値によるイベント別の検出性能の評価を行った．F 値による評価では，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われたと判定している．

4.2 実験結果

音声認識の結果を表 1 に，食行動音認識の結果を表 2 に示す．

表 1 より，音声認識の観点では，WavLM を用いて構築した認識モデルは，HuBERT の音声+食行動音でチューニングされた認識モデルに比べて優れた認識結果となった．これより，大規模事前学習モデルを WavLM にしたことで音声認識の観点では認識性能が向上したことが確認できた．一方で，音声のみでチューニングされたモデルと音声+食行動音でチューニングされたモデルにおける CER の差については，HuBERT では 2.9%だったのに対して，WavLM では 2.5%，拡張手法を適用したモデルでは 2.7%となった．この結果から，WavLM で認識モデルを構築したことにより 0.2～0.4%の改善が見られたが，食行動音との同時学習による音声認識の性能の低下という課題は，WavLM を用いても大きく改善することができないことが分かった．

表 2 より，食行動音認識の観点では，音声認識と同様に WavLM を用いることで認識率の改善が見られた．また，10 倍に拡張した食行動音でチューニングされた WavLM のモデルでは，咀嚼の認識率が LSTM-CTC の食事音認識モデルと同じ F 値となった．食行動音を 10 倍に拡張することによっても認識性能が向上することが確認できた．しかし，音声+食行動音でチューニングされた WavLM のモデルや，音声+10 倍拡張食行動音でチューニングされた WavLM のモデルに関しては，まだ食行動音認識専用に構築された LSTM-CTC モデルの認識性能には及ばないことも分かった．

表 1 音声認識の結果

ベースモデル	学習データ	チューニングデータ	CER(%)↓
HuBERT	LibriLight(60k hr)	音声(5 hr)	13.8
WavLM	Mix(94k hr)	音声(5 hr)	**10.9**
HuBERT	LibriLight(60k hr)	音声+食行動音(5+2.3 hr)	16.7
WavLM	Mix(94k hr)	音声+食行動音(5+2.3 hr)	**13.4**
WavLM	Mix(94k hr)	音声+10倍拡張食行動音(5+23 hr)	**13.6**

表 2 食行動音認識の結果

ベースモデル	学習データ	チューニングデータ	F値(イベント別)↑	
			咀嚼	嚥下
LSTM-CTC	食行動音(2.3 hr)	-	0.70	0.51
HuBERT	LibriLight(60k hr)	食行動音(2.3 hr)	0.49	0.20
WavLM	Mix(94k hr)	食行動音(2.3 hr)	**0.60**	**0.36**
WavLM	Mix(94k hr)	10倍拡張食行動音(23 hr)	**0.70**	**0.40**
HuBERT	LibriLight(60k hr)	音声+食行動音(5+2.3 hr)	0.31	0.18
WavLM	Mix(94k hr)	音声+食行動音(5+2.3 hr)	**0.56**	**0.26**
WavLM	Mix(94k hr)	音声+10倍拡張食行動音(5+23 hr)	**0.63**	**0.36**

5. おわりに

今回の実験では，WavLM 及び N-gram に基づくデータ拡張手法を用いた音声と食行動音の同時認識モデルを構築した．WavLM を用いることで音声認識及び食行動音認識のどちらにおいても精度改善が確認できた．また，食行動音のデータ拡張によって食行動音認識では精度改善が見られたが，音声認識では認識性能が低下してしまうことが分かった．

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，”wav2vec2.0 の事前学習モデルを用いた咽喉マイク音声認識”，情報処理学会第 84 回全国大会講演論文集，(2022)

[2] 小祝和寛　ほか，”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”，情報処理学会第 85 回全国大会講演論文集，(2023).

[3] Sanyuan Chen, et al. “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”, arXiv preprint arXiv:2110.13900, (2022).

[4] Akihiro Nakamura, et al. “Automatic Detection of Chewing and Swallowing”,Sensors 2021, Vol. 21 Issue 10, 3378, (2021).