事前学習済み深層学習モデルを用いた音声による認知症自動検出に関する日英比較

福田悠人[1] 西村良太[2] 吉田稔[1] 松本和幸[1]
[1]徳島大学 理工学部 理工学科 知能情報コース
[2]豊橋技術科学大学 情報・知能工学系
c612101269@tokushima-u.ac.jp nishimura.ryota.tz@tut.jp
mino@is.tokushima-u.ac.jp matumoto@is.tokushima-u.ac.jp

概要

現在，高齢化社会の進行に伴い認知症患者が増大している．医療機関での受診はコストや時間がかかる上，早期発見が困難であるという課題がある．本研究では，コストの高い書き起こしテキストを必要とせず，音声波形のみから手軽かつ迅速に認知症傾向を判定するモデルの構築を目的とする．自己教師あり学習モデルである HuBERT および wav2vec 2.0 を特徴抽出器として採用し，日本語および英語のデータセットを用いて分類精度の比較検証を行った．実験の結果，日本語データにおいて HuBERT を用いたモデルが 73.0%の精度を達成し，先行研究の機械学習ベースモデルを上回る性能を示した．

1 はじめに

高齢化社会の進行に伴い，認知症の患者数が増大している．しかし，認知症に対しては課題が多く存在している．医療機関での診察には多くの費用と時間がかかる上，専門家であっても症状が軽度な場合は発見が困難である．また，高齢者自身や家族が症状を見過ごしてしまい，そのまま重症化してしまうというケースも少なくない．現在，音声データから認知症を検出する試みは数多くなされているが，多くの手法では音声波形から抽出した音響的特徴と組み合わせて，手書きによる書き起こしテキストを用いた言語的特徴が利用されている．しかし，手書きの書き起こし作業はコストが高く高齢者が簡単に使用することが難しくなってしまう．そこで，本研究では書き起こしを必要とせず，音声波形のみから手軽で迅速な認知症傾向判定モデルの構築を目的とする．また，深層学習を用いた自己教師あり事前学習済みモデルである HuBERT および wav2vec 2.0 を特徴抽出器として採用し，日本語および英語のデータセットを用いて分類精度の比較検証を行う．

2 提案手法

本研究で構築したモデルは，生の音声波形を入力とし，特徴抽出層，プーリング層，全結合層から構成される．モデルに入力される音声データはスライド幅 2 秒で 5 秒ごとのチャンクに分割して処理を行う．特徴抽出層には事前学習済みモデルである HuBERT，および wav2vec 2.0 を用いる．抽出された特徴量は時間方向に平均プーリングされ，その後全結合層にて認知症傾向あり・なしの二値分類を行う．言語情報に依存せず，音響的特徴のみを用いることで，多言語への拡張性をもたせることができると考える．

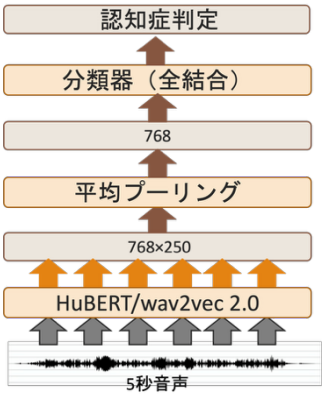

図 1 モデル構造

3 使用するデータと前処理

日本語コーパスとして「GSK2018-A 対照群付き高齢者コーパス」を用いた．このコーパスには 60～79 歳の高齢者 60 名と 20～59 歳の非高齢者 20 名の音声が存在している．自然文読み上げやイラスト描

写などの課題に対する回答音声が含まれており，話者ごとの MMSE テストの結果が付与されている. 後の実験ではこのテストの結果が 27 点以下の話者を認知症傾向者群に分類し，非認知症話者をまとめた非認知症話者群とともに使用した．英語コーパスとしては「INTERSPEECH 2020 ADRESS チャレンジ」内の cookie 盗難についての音声データを使用した．英語コーパスには話者ごとの認知障害診断結果が付与されており，認知症と診断された話者 194 名と非認知症話者 99 名の音声データが含まれている．それをもとにアルツハイマー型認知症と診断されている話者を認知症傾向者群に分類し，こちらも非認知症話者群とともに実験に使用した．英語コーパスの音声データには対象者以外の音声が含まれていたためコーパスに付属していたタイムスタンプ情報を利用して被験者の発話区間のみを抽出，結合した．さらに，音量のばらつきも大きかったため，極端に音量が小さい(平均音量−47db 以下)，あるいは大きい(平均音量−15db 以上)データを除去するフィルタリング処理を実施した.

4 実験・結果

4.1 日本語データにおける評価

日本語データを使用し，特徴抽出機に HuBERT(rinna/Japanese-hubert-base)を用いたモデルで認知症傾向者群と健常者群のデータ数比率を 1 : 2 に調節し学習を行った．実験の結果，学習に使用した話者(既知話者)の分類精度が 95.2%, 学習に使用していない未知話者の分類精度は 62.5%であった．認知症傾向者群と健常者群の比率を 1 : 1 に変更し再度実験を行うと, 既知話者対する分類精度は 98.3%, 未知話者に対しては 73.0%を達成した．これは，音響特徴量のみを用いた機械学習ベースの先行研究の精度を上回る結果である．一方，wav2vec 2.0(reazon-research/japanese-wav2vec2-base) を用いた場合, 未知話者に対する精度は 60.0%にとどまり，本実験では HuBERT の方が高い分類性能を示した.

表 1 ベースラインとの認知症識別精度比較

	Acc.
ベースライン	.708
HuBERT	.730
wav2vec 2.0	.600

4.2 英語データにおける評価・日本語との比較

英語データを用いた実験では，HuBERT (facebook/hubert-base-ls960)を用いた場合の未知話者精度は 61.7%となり，日本語データと比較して精度が低下した.

表 2 日本語・英語識別精度比較

	Acc.	F1	AUC
JP HuBERT 既知話者	.983	.967	.996
未知話者	.730	.759	.756
EN HuBERT 既知話者	.812	.812	.917
未知話者	.617	.577	.704

5 考察

データ数の均衡化を行うことで, 機械学習ベースの先行研究と比較しても高い精度が得られる可能性が示唆された．HuBERT を使用したモデルと wac2vec2.0 を使用したモデルで精度に大きな差が生まれたのは rinna モデルと reazon-research モデルで事前学習モデルに使用されている音声に差が存在していることが大きく影響していると考えられる．日本語データと英語データでの結果に大きな差が生まれた原因としては使用した 2 言語の音声データの間に音質の差があったことが考えられる．また，英語データの前処理の際にタイムスタンプを使用したため発話間の言い淀みや無音区間がなくなってしまったことも原因の一つとして考えられる.

6 まとめと今後の予定

本研究では，HuBERT を用いた音声波形のみによる認知症判定モデルを構築し，日本語データにおいて先行研究を上回る精度を確認した．今後は，特徴抽出に使用するモデルを多言語モデルに統一した場合における単一言語ごとの学習及び両言語の同時学習の比較を行う．また，英語データにおける前処理手法の改善とともに言い淀みや無音時間といった時間特徴の追加，および学習率等のハイパーパラメータの最適化を進め，さらなる高精度化を目指す.

謝辞

本研究は令和 7 年度自賠責運用益拠出事業により行われました．深く感謝いたします．

参考文献

[1] Nicholas Cummins, Yilin Pan, Zhao Ren, et al., “A Comparison of Acoustic and Linguistics Methodologies for Alzheimer’s Dementia Recognition”, Interspeech, 2020.

[2] Edward L. Campbell, Laura Docio-Fernandez, et al., "Alzheimer's Dementia Detection from Audio and Language Modalities in Spontaneous Speech", IberSPEECH, 2021.

[3] Nicholas Cummins, Yilin Pan, Zhao Ren, et al., “ Exploring multi-task learning and data augmentation in dementia detection with self-supervised pretrained models”, Interspeech, 2023.

[4] Benjamin Barrera-Altuna, Benjamin Barrera-Altuna , Daeun Lee , Zaima Zarnaz , Jinyoung Han, Seungbae Kim"The Interspeech 2024 TAUKADIAL Challenge: Multilingual Mild Cognitive Impairment Detection with Multimodal Approach", Interspeech 2024, pp. 967-971, 2024.