E-012

HuBERT を用いた音声感情認識における Fine-tuning の効果
Effects of Fine-tuning HuBERT on Speech Emotion Recognition

久下 忍＊　柘植 覚†　西田 昌史‡　堀内 靖雄＊　黒岩 眞吾＊
Shinobu Kuge　Satoru Tsuge　Masafumi Nishida　Yasuo Horiuchi　Shingo Kuroiwa

1. はじめに

音声感情認識は，話者が音声に込めた感情を推定する技術であり，ヒューマンロボットインタラクションやコールセンターなど，さまざまな場面での応用が期待されている．近年では，大量のラベルなし音声データから音響特徴抽出器として学習された自己教師あり学習モデルの音声感情認識に対する有効性が報告されている[1]．さらに，事前学習済みの自己教師あり学習モデルに対して，感情ラベル付き音声データを用いた Fine-tuning を行うことで，精度向上が可能であることも示されている [2]．しかし，これまでの研究の多くは英語などの非日本語データを対象としており，日本語の感情音声コーパスを用いた Fine-tuning に関する研究は限られている[3].

そこで本研究では，自己教師あり学習モデルの 1 つである HuBERT[4]に対して，日本語感情音声コーパス JTES[5]を用いた Fine-tuning を行う．なお，Fine-tuning を行う際には，Transformer 内部の学習層数を変更し，認識精度を比較する．さらに，認識対象の発話文の長さが認識精度に与える影響についても検証を行う．

2. 提案手法

本研究では，日本語音声を対象とした感情認識モデルを構築するため，事前学習済み HuBERT に対して日本語感情音声コーパスを用いた Fine-tuning を行う．HuBERT には，rinna 社が日本語音声コーパス ReazonSpeech v1 を用いて事前学習を行った rinna/japanese-hubert-base（95M パラメータ）[6]を使用する．

図 1 に使用するモデルの構造を示す．このモデルは，Transformers ライブラリにおける HubertForSequence-Classification の構造に準拠している．まず，入力音声は CNN および Transformer（全 12 層の Multi-Head Attention 層+Frame-Wise Linear 層）から構成される HuBERT に入力され，時系列（フレーム方向）に沿った音響特徴埋め込みが抽出される．瀧澤大吾らの研究[3]では，Transformer の各層からの出力に対しての重み付き和を最終的な埋め込みとしているのに対し，本研究では最終層の出力のみを埋め込みとして用いる．次に，得られたフレームレベルの埋め込みに対して，Frame-Wise Linear 層を適用することで次元数を削減し，さらにフレーム方向に対して平均プーリングを行う．これにより，発話レベルの 256 次元ベクトルが得られる．この発話レベルのベクトルに対して，Linear 層および Softmax 関数を適用し，出力の次元を感情クラス数（本研究では 4 クラス）まで削減する．なお，Fine-tuning を行う際には，HuBERT の全体構造のうち，CNN および Transformer 内部の入力に近いm層のパラメータを固定し，残りの$n = (12 - m)$層のパラメータのみを更新対象とする．これにより，事前学習で得られた汎用的な音響特徴表現を保持しつつ，感情認識に特化した学習を実現する．実験では，nを 0 から 12 の範囲で変化させることで，事前学習済み音響特徴表現の活用度合いと，Fine-tuning によるタスク適応効果の関係について比較を行う．

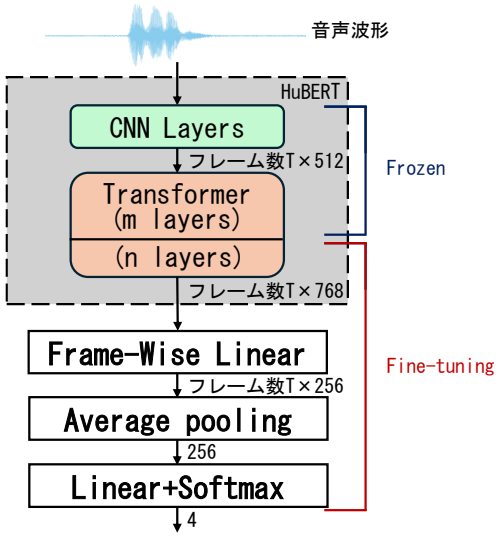

図 1　HuBERT ベースのモデル構造

3. 評価実験

3.1 実験条件

実験では，日本語感情音声コーパスである JTES（Japanese Twitter-based Emotional Speech corpus）[5]を用いた．JTES は，話者 100 名（男女各 50 名）が 4 感情（喜び・怒り・悲しみ・平常）で発話した音声を収録したコーパスである．各話者は，異なる音韻や韻律のバランスを考慮して選出された各感情 50 文を発話している．収録は，サンプリング周波数 48kHz，量子化精度 16bit で，防音室にて行われた．実験では，音声を 16kHz にダウンサンプリングして使用した．

Fine-tuning に用いる学習データセットには，話者 80 名(男女各 40 名）×30 文×4 感情の計 9,600 発話を使用し，検証・評価データセットには，それぞれ話者 10 名(男女各 5 名)ずつの 10 文×4 感情の各 400 発話を使用した．各データセット間では，話者および発話文の重複は排除し，話者およびテキスト非依存となるように分割を行った．また，発話文の長さにより各感情クラスに対する認識精度に偏りが生じる可能性を考慮し，評価データセットにおいては，すべての感情クラスで平仮名文字数が同一の発話文を選定した．検証データセットについても同様の条件を適用した

＊千葉大学 Chiba University
†大同大学 Daido University
‡静岡大学 Shizuoka University

が，4 感情すべてで文字数が同一の発話文が存在しない場合には，前後 1〜2 文字の差異を許容して選定を行った．さらに，各データセット間で発話文の文字数に偏りが生じないよう，短文から長文までをバランスよく含めるよう配慮した．なお，本研究では各感情クラス間で発話数に偏りがないため，評価指標には感情ごとの平均を取らずに全体での認識精度を用いた．

Transformer の凍結条件を変化させた各モデルの学習におけるハイパーパラメータ（learning rate，batch size など）は，自動最適化フレームワークである Optuna を用いて決定し，エポック数は 30 とした．

3.2 実験結果

まず，Transformer 内部の出力に近いn層を Fine-tuning の対象とし，nを 0 から 12 の範囲で変化させたモデルの感情認識結果を図 2 に示す．図 2 より，2 層までを Fine-tuning した場合と，3 層以上を Fine-tuning した場合とで，認識精度に大きな差が生じていることが分かる．一方で，HuBERT を Fine-tuning せずに音響特徴抽出器として用いた場合（$n = 0$）の精度は，2 層までを Fine-tuning した場合に比べて高くなっており，少数の層のみを Fine-tuning することで，精度の低下を招く可能性があることが示唆された．さらに，3 層以上を Fine-tuning した場合，有意な差とは言えないが，8 層を学習対象とした際に最も高い精度となった．このときの認識精度は，大規模な自己教師あり学習モデル(317M パラメータ)を特徴抽出器として用いた Atmaja らの研究[1]で報告された，HuBERT Large および WavLM Large の結果をそれぞれ 1.7%，0.2%上回った．なお，Transformer 内部の全層を Fine-tuning した場合（$n = 12$）には，精度がピーク時（$n = 8$）に比べて低下することが確認された．以上より，HuBERT の Transformer 層を Fine-tuning することで，より高い認識精度が得られることが示された．

続いて，最も高い認識精度を示した$n = 8$における混同行列および，発話文の文字数ごとの感情認識結果をそれぞれ図 3，図 4 に示す．図 3 からは，「怒り」や「平常」の感情が，「喜び」に誤って認識される傾向にあることが示された．一方で，4 感情の中で最も認識しやすい感情は「悲しみ」であった．さらに，図 4 からは，発話文の文字数が長くなるほど認識精度が向上する傾向が見られる．17 文字以上の発声のみで精度を算出すると 90.5%となった．このことから，短い発話文に対するモデルの頑健性向上が，感情認識精度全体の向上に向けた課題であるといえる．

4. おわりに

本稿では，日本語音声における感情認識モデルの構築を目的として，自己教師あり学習モデル HuBERT に対して，日本語感情音声コーパス JTES を用いた Fine-tuning を行った． 実験の結果，Transformer 内部の 12 層中 8 層を学習対象としたときに，最も高い認識精度が観測されたものの，その精度は 78.5%にとどまった．しかし，発話文の文字数ごとの認識精度を分析したところ，17 文字以上では 90.5%の精度が得られており，短い発話文に対する精度が全体の精度を押し下げていることが示唆された．このことから，今後は短い発話文に対する頑健性を高める手法の検討を行う．

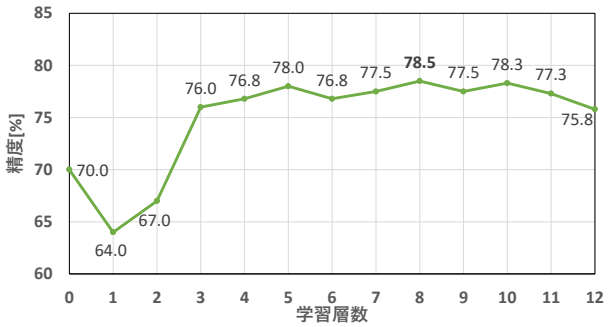

図 2　学習層数nと感情認識精度の関係

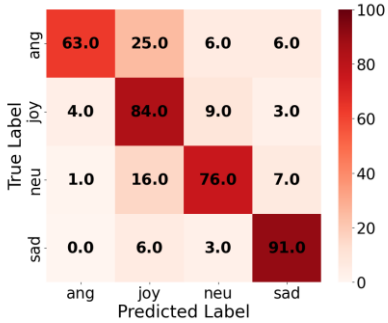

図 3　混同行列[%]（$n = 8$）

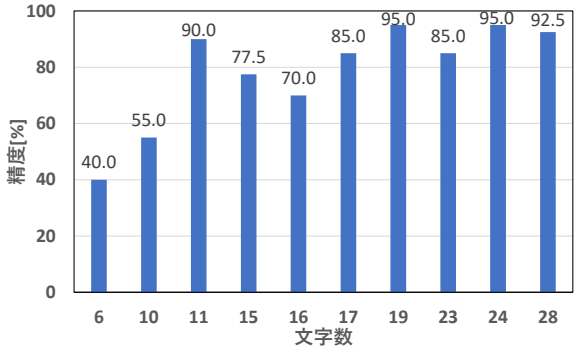

図 4　文字数ごとの感情認識精度（$n = 8$）

謝辞

本研究は，JSPS 科研費 24K07957，23K11165，24K14988，25K15125 の助成を受けたものです．

参考文献

[1] B.T. Atmaja, A. Sasou, “Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition,” IEEE Access, Vol.10, pp.124396–124407, 2022.
[2] Edmilson Morais et al, “Speech Emotion Recognition using Self-Supervised Features,” ICASSP, pp. 6922–6926, 2022.
[3] 瀧澤大吾ら, “日本語音声感情認識のための自己教師あり学習モデルの検討,” 日本音響学会第 150 回講演論文集, pp.1541–1544, 2023.
[4] W.-N. Hsu et al, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” IEEE/ACM Trans. on Audio, Speech and Language Processing, Vol.29, pp.3451–3460, 2021.
[5] E. Takeishi et al, “Construction and Analysis of Phonetically and Prosodically Balanced Emotional Speech Database,” Oriental COCOSDA, pp.16–21, 2016.
[6] Y. Hono et al, “https://huggingface.co/rinna/japanese-hubert-base,” 2023.