J-015

マルチモーダルデータにおけるうつ状態検出のための音声特徴の抽出と比較
Extraction and Comparison of Acoustic Features for Depression State Detection Using Multimodal Data

岡田 大和[1)] 高鍋 俊樹[1)] 松本 和幸[1)] 柏原 巧太郎[2)] 木内 敬太[3)]
Yamato Okada Toshiki Takanabe Kazuyuki Matsumoto Kotaro Kashihara Keita Kiuchi

梅原 英裕[1)] 中瀧 理仁[1)] 沼田 周助[1)] 吉田 稔[1)] 康 鑫[1)]
Hidehiro Umehara Masahito Nakataki Shusuke Numata Minoru Yoshida Xin Kang

1 はじめに

うつ病は世界保健機関（WHO）によって，全世界で約 3 億人以上が影響を受けているとされる深刻な精神疾患であり，心身の健康や社会的・経済的な機能に重大な影響を及ぼす [1]．特に，うつ病による生産性低下や自殺リスクの増加は，個人にとっての苦痛のみならず，社会全体にも大きな損失をもたらす．そのため，早期発見と適切な介入が重要であることは言うまでもない．

現在のうつ病診断は主に精神科医による問診と心理検査に依存しているが，このプロセスには主観性や臨床経験への依存性，さらには医療リソースの制限といった課題が伴う．こうした背景から，人工知能（AI）技術を活用した客観的かつ再現可能な診断支援システムの開発が求められている．

本研究では，うつ病の診断補助を目的として，発話音声に着目し，音響的特徴量を用いた分類モデルの構築とその性能評価を行う．特に，入力特徴量に対する寄与分析を通じて，日本語におけるうつ病検出においてどのような音声的特徴が重要であるかを明らかにし，今後の臨床応用や多言語対応への足がかりとする．

2 関連研究

音声によるうつ病検出に関する研究は，近年のディープラーニング技術の進展により飛躍的に進展している．Han ら [2] は，公開コーパスである DAIC-WOZ を用いて，音声特徴量と畳み込みニューラルネットワーク（CNN）を組み合わせたモデルを構築し，高精度なうつ病検出を実現した．この研究は，音声の時系列情報と深層学習の組み合わせが有効であることを示す好例である．

また，感情認識の分野で広く利用されている特徴量セットとして，Eyben らによって提案された GeMAPS およびその拡張版 eGeMAPSv02 がある [3]．これらの特徴量セットは，感情や精神状態を反映しやすい音声的指標を体系的に選定したものであり，信頼性と再現性の高い解析が可能である．

しかし，これらの先行研究の多くは英語を対象としており，日本語におけるうつ病検出に関する研究は限られている．特に，日本語話者の文化的・言語的特性が音声的特徴にどのような影響を与えるかについての知見は乏しい．本研究では，この点に着目し，日本語データにおける特徴量の妥当性と分類性能について詳細に検討を行う．

1） 徳島大学 Tokushima University
2） ワークスアプリケーションズ Works Applications
3） 労働者健康安全機構 Japan Organization of Occupational Health and Safety

3 手法

3.1 データセットと収集方法

本研究では，日本語と英語の 2 種類の面談音声データを用いた．1 つは徳島大学が独自に収集した日本語音声データセットであり，参加者 101 名（うちうつ病と診断された者 22 名）から構成される．このデータは精神科医との連携のもとで収集されており，臨床的妥当性のある高品質な音声記録が含まれている．もう 1 つは DAIC-WOZ という米国発の公開データセットであり，189 名の面談音声と精神状態ラベルを含んでいる．

独自データの収集には，Kashihara ら [4][5] の研究に基づき，アバターや音声変換技術を取り入れた面談環境を整備し，インタビュアーの違いによる参加者へ与える影響を最小化する工夫を行った．Zoom を用いて対話を実施し，音声・映像・心拍・テキストデータを同時収録した．これにより，多角的なデータ解析が可能なマルチモーダルデータセットが構築された．

精神状態の評価には PHQ-9 を使用し，スコアの閾値を 10 以上とすることで「うつ」と「非うつ」の 2 値ラベルを付与した．参加者は事前に研究内容の説明を受け，同意を得たうえで自己記入形式のアンケートに回答し，倫理的配慮にも十分留意した．

3.2 特徴量抽出と前処理

音声データからの特徴量抽出においては，まずはフレームごとの特徴を抽出するために、librosa を用いて 1 秒ごとにフレーム分割を行い，openSMILE（ver3.0）で eGeMAPSv02 に含まれる 88 次元の低レベル記述子（LLD）を抽出した．これには，ピッチ（F0），声の振幅，フォルマント，ジッター，シマー，スペクトル重心，メル周波数ケプストラム係数（MFCC）などが含まれる．これらの特徴量は，声の高さ・不安定さ・共鳴特性など，うつ症状と関連があるとされる音響的要素を網羅しており，多角的な分析が可能である．

また、被験者間の発話長の差異に対応するため，系列長の最大値に合わせてゼロパディングを施した．これにより，可変長系列を固定長テンソルとして扱うことが可能となる．さらに，各特徴量は z スコアによる正規化を行い，異なるスケールを持つ特徴間のバランスを調整し，学習の安定性と収束性を高めた．

3.3 分類モデル構成

提案モデルは，2 層の畳み込み層（各 32 チャネル，カーネルサイズ 3），ReLU 活性化関数，バッチ正規化，2 層の最大プーリング層を組み合わせた CNN 構造である．Flatten 後には，中間層 64 ユニットの全結合層を 2 層通過させ，抽出された音響特徴の高次抽象化を行う．最終層には softmax 関数を用いてうつ／非うつの 2 クラス分類を行う構成とした．

最適化には Adam を用い，損失関数にはクロスエントロピーを採用した．データにおけるクラスの不均衡（うつ患者が少数派）に対応するため，クラスごとに逆数重みを付けた重み付き損失関数を用いた．さらに，検証データに対する性能の悪化を検知し，過学習を抑制するために Early Stopping を導入した．学習は最大 50 エポックで行い，3 エポック連続で性能改善が見られなければ停止した．

4 実験と評価

4.1 分類性能の評価

DAIC-WOZ と日本語音声データセットに対してクロスバリデーションを実施した．結果として，表 1 のようになった．

表 1 各データセットにおける分類結果（%）

指標	DAIC-WOZ		日本語データ	
	非うつ	うつ	非うつ	うつ
Precision	77	50	79	29
Recall	82	43	69	40
F1 値	80	46	73	33
Accuracy	70		62	

4.2 寄与特徴量の分析

モデルの説明性を向上させるため，Integrated Gradients [6] を用いて入力特徴量に対する勾配ベースの寄与分析を行った．この手法は，ニューラルネットワークにおける各入力特徴の予測への寄与度を定量的に評価できるものであり，特に深層学習モデルのブラックボックス性を軽減する手法として注目されている．その結果，MFCC，基本周波数（F0），スペクトルフラックス，およびフォルマント周波数などの特徴量が，うつ状態の判別において強く寄与していることが確認された．これらの特徴量は，声の高さ，音の滑らかさ，母音の構造的性質など，うつ状態に伴う音声的変化と一致している．

また，日本語音声データセットでは，英語話者を対象とした DAIC-WOZ と比較して，スペクトル関係の特徴量の寄与率が相対的に高くなる傾向が見られた．特に，スペクトル中心やスペクトルスプレッドといった特徴量が高い重要度を示した．これは，日本語に特有の音韻的性質や発話スタイルの違いが，うつ状態に現れる音声的特徴に影響を与えている可能性を示唆している．すなわち，言語的背景によってうつ病の音声的兆候が異なる可能性があり，言語依存性を考慮したうつ検出モデルの設計が今後の課題となる．

5 考察

本研究では，音響特徴量を用いたうつ病検出において，英語話者（DAIC-WOZ）で比較的高い分類性能が得られた一方，日本語話者では精度がやや低下した．これは，データ数やクラス不均衡，文化的な発話傾向の違いが影響していると考えられる．

また，言語的・文化的背景の違いもうつ病傾向の音声表現に影響していると考えられる．日本語話者は一般に抑揚が少なく，感情を声に出すことを避ける傾向があり，健常者とうつ病患者の音響的な差が欧米言語より小さい可能性がある．

寄与分析においては，F0，スペクトルフラックス，MFCC といった特徴が健常者とうつ病患者の判別に寄与することが示された．これらは過去の研究とも一致しており，音響的な平坦さやピッチの低下などがうつ病の指標として有効であることを支持している．

6 今後の課題

今後は，より高次の時間的依存関係を捉えることができる Transformer アーキテクチャや，計算効率と精度の両立を図ることができる時間畳み込みネットワーク（TCN）などの導入が有望である．これにより，音声信号の微細な変化や文脈的な連続性をより精緻に捉えることが可能となる．

また，本研究では音声特徴量のみに焦点を当てたが，現実の面談環境では音声以外にも多くの情報が存在する．たとえば，表情や視線，姿勢といった非言語的行動，さらには発話内容そのものも，うつ状態の判別において重要な手がかりとなりうる．これらを統合的に扱うマルチモーダル解析は，より精度の高いモデル構築に向けて不可欠な方向性である．

さらに，臨床現場への応用を視野に入れる場合，モデルの説明性（Explainability）や頑健性（Robustness），そしてユーザーとのインタラクション設計にも十分な配慮が必要となる．これらの課題を総合的に検討することで，実運用に耐える AI ベースのうつ病診断補助システムの実現が期待される．

謝辞

本研究は，令和 6 年度徳島大学ものづくり未来共創機構実証研究推進プロジェクトおよび，公益財団法人 JKA 令和 6 年度開発研究補助事業および，国立研究開発法人科学技術振興機構（JST）大学発新産業創出基金事業スタートアップ・エコシステム共創プログラム「PSI・GAP ファンド支援プログラム（ステップ 1）」により実施されました．深く謝意を表します．

参考文献

[1] 世界保健機構（WHO），”Depression and Other Common Mental Disorders”.

[2] Han, J. et al., ”Speech-based depression detection using deep learning,” *Proc. Interspeech*, 2018.

[3] Eyben, F. et al., ”The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing”, *IEEE TAC*, 2016.

[4] Kashihara, K. et al., ”Constructing multimodal counseling dataset for depressive states and feature analysis”, Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval, 2024.

[5] 柏原 功太郎，髙鍋 俊樹，木内 敬太，梅原 英裕，入澤 航史，中瀧 理仁，沼田 周助，康 鑫，吉田 稔，松本 和幸：早期うつ状態検出のためのマルチモーダル対話データセットに基づくうつ状態検出モデルの性能評価，言語処理学会第 31 回年次大会発表論文集，1393-1397，2025 年 3 月．

[6] Sundararajan, M. et al., ”Axiomatic Attribution for Deep Networks,” *Proc. ICML*, 2017.