卒 業 論 文 概 要 書

Summary of Bachelor's Thesis

Date of submission: 01/30/2024 (MM/DD/YYYY)

学科名 Department	情報通信	氏名 Name	細郷壮希	指 導 教 員 Advisor	渡辺 裕 印
研究指導名 Research guidance	オーディオビジュアル 情報処理研究	学籍番号 Student ID number	1W202146-0		
研究題目 Title	Bi-LSTM 分類器を用いた音声・テキスト・表情によるマルチモーダル感情推定の精度改善 Improving the Accuracy of Multimodal Emotion Estimation by Speech, Text and Facial Expression Using Bi-LSTM Classifier				

1. まえがき

近年，デジタルコミュニケーションの進化に伴い，さまざまなコンテンツ上での感情表現の機会が飛躍的に増加している．コンピュータと人間のインタラクションを円滑にするためには，複数のモーダルの統合により人間の感情をより正確に推定することを目指すマルチモーダル感情推定技術が不可欠である．この技術は自動運転の安全システム[1]や病気の治療促進[2]，表情をトレーニングするアプリケーションなどといった幅広い分野での応用が期待されている．しかし，広く実用化を目指した場合，正確性やリアルタイム処理が大きな障壁となっている．

本論文では従来の音声・テキストによる感情推定手法[3]および音声・表情による感情推定手法[4]をもとに，より広範囲な時系列情報を捉えられる感情推定手法を提案する．新たなモーダルを加えるとともに双方向の時系列情報を付加することで，感情クラスの誤分類が改善されることを示す．

2. 従来手法とその問題点

2.1 音声・テキストによる感情推定手法

Long Short-Term Memory (LSTM)分類器を用いて音声・テキストにより感情を推定する手法が提案されている．音声とテキストの特徴ベクトルを連結したベクトルを LSTM ネットワークへの入力とする．LSTM ネットワークの出力をソフトマックス層に通すことで 6 感情 (憤怒，幸福，悲哀，恐怖，驚愕，中立) のクラス分類を実行する．LSTM の特性から長期記憶における情報の重要度を考慮した上で感情を分類することができる利点がある．

2.2 音声・表情による感情推定手法

Bidirectional Long Short-Term Memory (Bi-LSTM) 分類器を用いて音声・表情により感情を推定する手法が提案されている．この手法では，まず表情の特徴ベクトルを Bi-LSTM ネットワークへ入力する．同時に，音声の特徴ベクトルを，微調整した事前学習トランスフォーマモデルと MLP へ入力する．その後，それぞれから得られた各感情クラスに対する確率分布を結合し，8 感情（憤怒，幸福，悲哀，恐怖，驚愕，中立，平静，嫌悪）のクラス分類を実行する．

Bi-LSTM ネットワークでは LSTM ネットワークに対し追加の LSTM レイヤーを導入することで過去から未来への情報処理だけでなく，未来から過去への情報処理も可能であるため，ある時点の感情をその前後の情報を加味した上で推定することができる．

2.3 問題点

従来手法のうち 2.1 節の手法では，音声とテキストの融合により音声単体の場合と比較して，幸福クラスや恐怖クラス，驚愕クラスの誤分類が大きく改善される．一方で，2.2 節の手法では，音声と表情の融合により音声単体の場合と比較して，悲哀クラスや幸福クラス，憤怒クラスの誤分類が少ないという結果が得られている．

すなわち，従来の感情推定のアプローチでは個々の手法において特定の感情の識別精度が低いという欠点がある．

3. 提案手法

二つの従来手法を組み合わせることで，分離が容易でなかったクラスの推定性能を改善し，誤分類を削減できると考えられる．そこで提案手法では，2.1 節のモデルに対し入力として新たに表情を追加し，2.2 節の Bi-LSTM ネットワークを適用する．

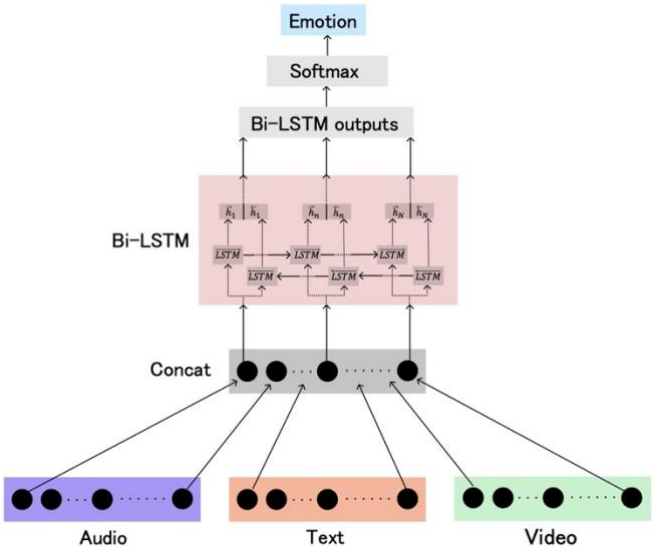

図 1 提案手法のモデル構造

提案手法のモデル構造を図 1 に示す．提案モデルでは，音声とテキスト，表情を連結したベクトルを Bi-LSTM ネットワークへ入力し，その出力をソフトマックス層に通すことで 6 感情のクラス分類を実行する．

4. 評価実験と結果

分類器の学習およびテストには IEMOCAP データセット[5]を用いる．損失関数にはクロスエントロピー損失を用い，最適化アルゴリズムとして Adam を採用する．

評価指標には従来手法 (2.1 節) との比較のため，正解率(accuracy)，適合率(Precision)，再現率(Recall)，F 値(F-measure)を用いる．従来手法と提案手法の評価結果を表 1 に示す．また，混同行列を図 2，図 3 に示す．

表 1 より，提案手法がすべての指標で従来手法を上回った．図 2，図 3 より従来手法において問題であった悲哀クラスとその他のクラスの混同が大きく改善された．また，ほとんどの感情で誤分類が改善され，精度の向上を確認できた．

表 1 各手法の評価結果

	Conventional	Proposed
Accuracy	59.8	**69.2**
Precision	62.3	**68.8**
Recall	59.3	**71.4**
F-measure	60.1	**69.3**

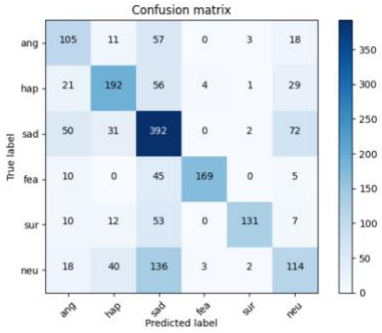

図 2 従来手法の混同行列

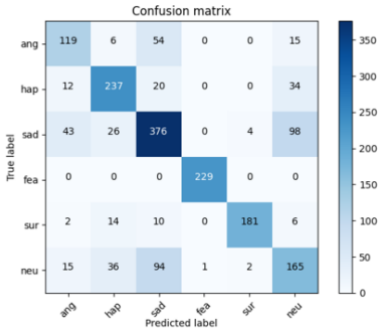

図 3 提案手法の混同行列

5. 結論

本研究では，従来の二つの感情推定手法をもとに，Bi-LSTM 分類器を用いた音声・テキスト・表情によるマルチモーダル感情推定手法を提案した．さらに，実験により提案手法の有効性を確かめた．

今後は，複数モーダルの融合方法やデータセットの処理方法と新たな特徴量の検討を行い，より効率的に感情を推定できるモデル構造を模索する．

参考文献

[1] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, R.W. Picard, “Driver Emotion Recognition for Intelligent Vehicles,” ACM Computing Survey, vol. 53, pp. 1-30, Jul. 2020.

[2] A.C. Nyquist, A.M. Luebbe, “An Emotion Recognition-Awareness Vulnerability Hypothesis for Depression in Adolescence: A Systematic Review,” Clinical Child and Family Psychology Review, vol.23, pp. 27-53, Mar. 2019.

[3] Gaurav. Sahu, “Multimodal Speech Emotion Recognition and Ambiguity Resolution,” CoRR, abs/1904.06022, Apr. 2019.

[4] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J.M. Montero, F. Fernández-Martínez, "A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset," Applied Sciences, vol. 12, no. 1, p. 327, Dec. 2022.

[5] C. Busso, M. Bulut, C-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, Nov. 2008.

2023 年度　卒業論文

Bi-LSTM 分類器を用いた音声・テキスト・表情による マルチモーダル感情推定の精度改善

Improving the Accuracy of Multimodal Emotion Estimation by Speech, Text and Facial E26xpression Using Bi-LSTM Classifier

指導教員　渡辺 裕　教授

提出日：2024 年 1 月 30 日

早稲田大学　基幹理工学部　情報通信学科

1W202146-0

細郷 壮希

目次

第１章　序論

１．１　研究背景

感情は，人間の日常生活におけるあらゆるコミュニケーションにおいて考慮すべき重要な要素である．感情を理解することで相手がどのように感じているかを把握することができ，より効率的なコミュニケーションや情報伝達が可能となる．一般に，感情は言葉のニュアンス，声のトーン，表情，身体言語など多様な手段を通じて表現されるが，それら単体では複雑な感情を表現し得ない．人間の感情は複数の要素を組み合わせて初めて正確に表現される．

近年，デジタルコミュニケーションの進化に伴い，ソーシャルネットワークサービスやオンラインプラットフォームをはじめさまざまなコンテンツ上で感情表現の機会が飛躍的に増加している．そのため，人間とコンピュータ間のインタラクションにおいて感情を正確に理解し反映する，すなわちユーザーの感情状態をリアルタイムで正確に把握し，それに適切に対応するシステムの需要が高まっている．そこで，注目されているのがマルチモーダル感情推定である．マルチモーダル感情推定は，音声や表情，テキストなどの複数のモーダルを統合し，より正確に人間の感情を推定することを目指している．

マルチモーダル感情推定は，自動運転の安全システム[1]や表情をトレーニングするアプリケーション，病気(うつ病[2,3]やパーキンソン病[4]など)の治療促進，教育分野でのインタラクティブな学習ツールなど多岐にわたる分野での応用が期待されている．しかし，広く実用化を目指した場合，感情推定の精度を飛躍的に高め，リアルタイムでの処理能力を向上させる必要がある．このように，多様な表現手段を正確に解釈し，統合することは容易ではないが，人間同士のコミュニケーション，さらには人間と機械とのインタラクションにおいて非常に重要な技術である．

１．２　関連研究と問題点，および研究目的

感情推定は深層学習の進展に伴い，長く注目されている研究分野である．とりわけモーダルの融合は推定精度を向上させる上で最も基本的な手法の一つである．しかし，人間の感情を判別する上で有用である音声，表情，テキストの中で一つ，二つを組み合わせるものは数多くあるが，三つ全てを用いるものはデータセットが不足していることから多くない．そこで本研究では従来の手法に対し，新たなモーダルを加えるとともに双方向の時系列情報を扱える分類器を導入したマルチモーダル感情推定手法を提案する．

１．３　本論文の構成

本論文の構成を以下に示す．

第１章　本章であり，本研究の背景，関連研究と問題点および研究目的について述べる．
第２章　本研究で用いる従来の感情推定手法および関連技術について述べる．
第３章　本研究の提案手法について述べる．
第４章　本研究における実験の方法，結果および考察について述べる．
第５章　結論と今後の課題について述べる．

第２章　　関連技術

２．１　　まえがき

本章では，従来の Long Short-Term Memory (LSTM)[5]分類器を用いた音声・テキストによる感情推定手法[6]および Bidirectional Long Short-Term Memory (Bi-LSTM)[7]分類器を用いた音声・表情による感情推定手法[8]について述べる．

２．２　　LSTM 分類器を用いた音声・テキストによる感情推定手法

LSTM 分類器を用いて音声・テキストにより感情を推定する手法が提案されている．音声とテキストの特徴ベクトルを連結したベクトルを LSTM ネットワークへの入力とする. LSTM ネットワークの出力をソフトマックス層に通すことで 6 感情（憤怒, 幸福, 悲哀，恐怖，驚愕，中立）のクラス分類を実行する．ソフトマックス層では入力された値を正規化して計算された各クラスに属する確率が出力される．従来手法[6]におけるモデル構造の概要を図２．１に示す.

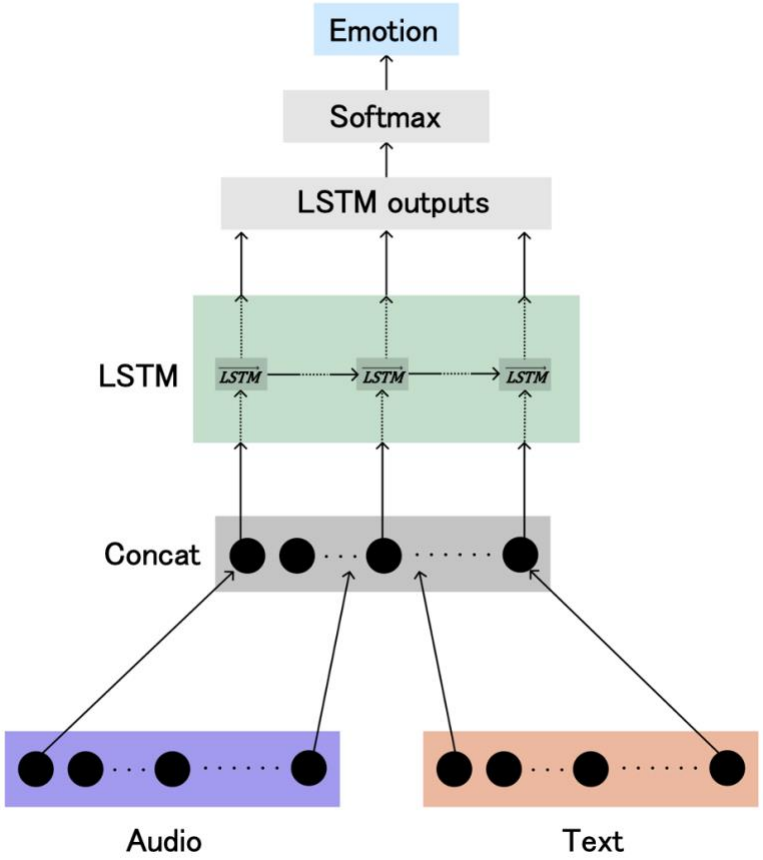

図　２．１　音声・テキストによる感情推定手法のモデル構造

LSTM は時系列データを扱うことができる Recurrent Neural Network (RNN)を改良したネットワークであり，長期記憶における情報の重要度を決定できるフィードバック接続を有する．この機構により，RNN において長いシーケンスを学習する際に初期の情報がネットワークを通じて薄れてしまうという勾配消失問題の解決が可能である．LSTM は新しい情報を追加するための入力ゲート，不要な情報を削除するための忘却ゲート，次の隠れ層への出力を決定するための出力ゲートの 3 種類のゲートを有しており，その構造を図２．２に示す．各時刻tで入力シーケンスx_tに対して過去の情報を蓄積した隠れ状態h_tとセル状態c_tを更新する一連の方程式は，式(2.1)-式(2.5) で表される．

$$f_t = \sigma_g(W_f x_t + U_f h_t + b_f) \tag{2.1}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{2.2}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{2.3}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{2.4}$$

$$h_t = o_t \cdot \sigma_h(c_t) \tag{2.5}$$

ここに，初期値は$c_0 = 0$，$h_0 = 0$，tはタイムステップ，・は行列の成分同士の積，x_tは LSTM ユニットへの入力ベクトル，f_tは忘却ゲートの活性化ベクトル，i_tは入力ゲートの活性化ベクトル，o_tは出力ゲートの活性化ベクトル，h_tは隠れ状態ベクトル，c_tはセル状態ベクトル，そしてW，U，bは重み行列とバイアス行列である．

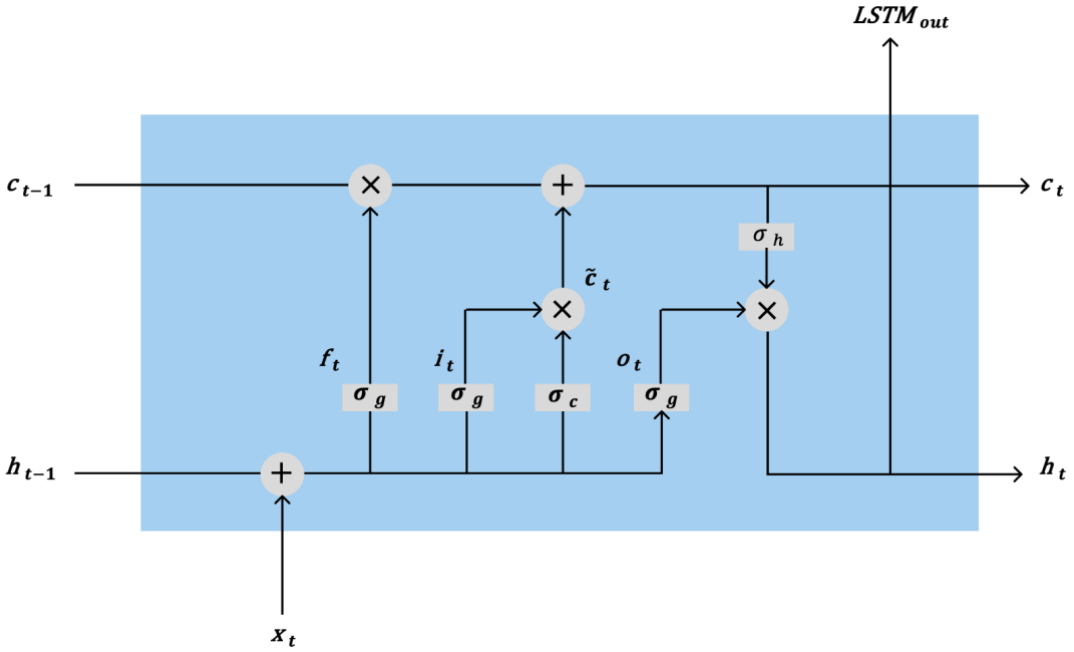

図 ２．２ LSTM ネットワークの概略図

２．３ Bi-LSTM 分類器を用いた音声・表情による感情推定手法

Bi-LSTM 分類器を用いて音声・表情により感情を推定する手法が提案されている．この手法では，まず表情の特徴ベクトルを Bi-LSTM ネットワークへ入力する．同時に，

音声の特徴ベクトルを，微調整した事前学習トランスフォーマモデルと多層パーセプトロン (MLP)へ入力する．その後，それぞれから得られた各感情クラスに対する確率分布である posterior ベクトルを結合し，多項ロジスティック回帰により 8 感情（憤怒，幸福，悲哀，恐怖，驚愕，中立，平静，嫌悪）のクラス分類を実行する．従来手法[8]におけるモデル構造の概要を図２.３に示す.

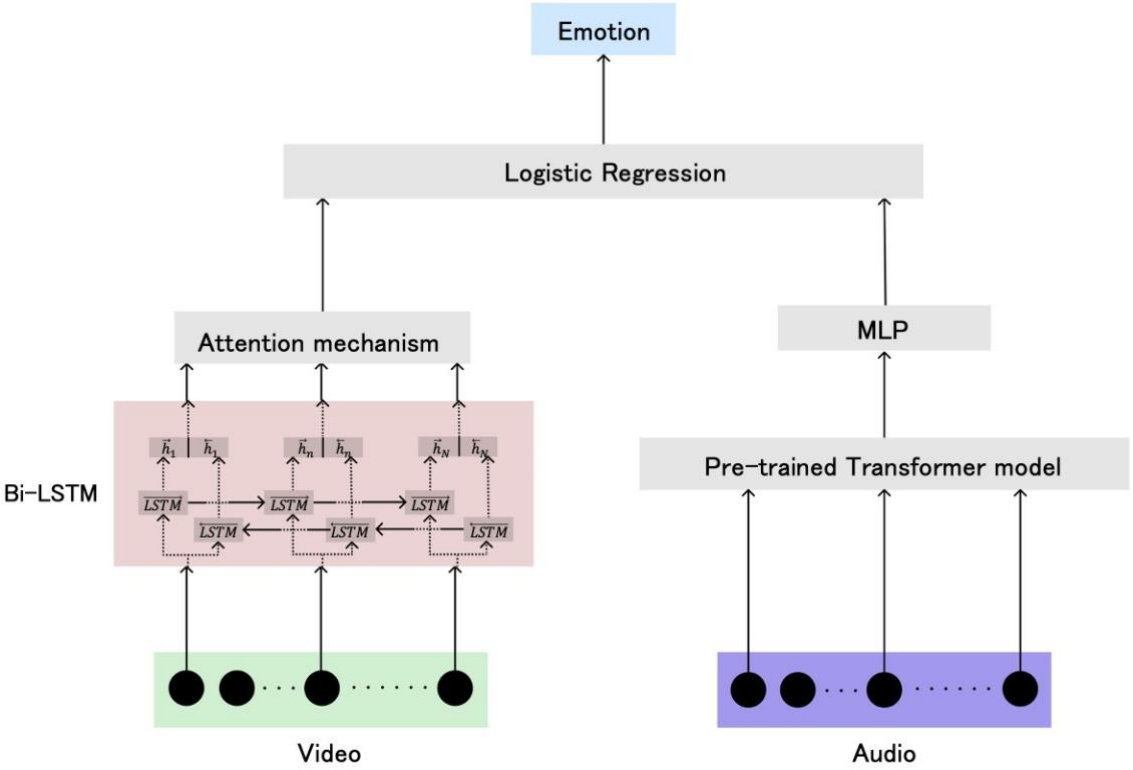

図 ２.３ 音声と表情による感情推定手法のモデル構造

Bi-LSTM は過去の重要な情報を長期間保持できる LSTM に加え，未来の情報も加味できるよう改良されたニューラルネットワークである．通常の LSTM が時系列データを前方向（過去から未来へ）に情報処理するのに対し，Bi-LSTM は追加の LSTM レイヤーを導入して後方向（未来から過去へ）に情報処理を行う．このネットワークはある時点での情報が前後の情報とどのように関連しているかモデルに把握させるのに役立つ．前方向 LSTM は式(2.6)-式(2.10)，後方向 LSTM は式(2.11)-式(2.15)，結合した隠れ状態ベクトルは式(2.16)で表される．

・Forward LSTM :

$$f_t^{(f)} = \sigma_g(W_f x_t + U_f h_{t-1}^{(f)} + b_f) \tag{2.6}$$

$$i_t^{(f)} = \sigma_g(W_i x_t + U_i h_{t-1}^{(f)} + b_i) \tag{2.7}$$

$$o_t^{(f)} = \sigma_g(W_o x_t + U_o h_{t-1}^{(f)} + b_o) \tag{2.8}$$

$$c_t^{(f)} = f_t^{(f)} \cdot c_{t-1}^{(f)} + i_t^{(f)} \cdot \sigma_c(W_c x_t + U_c h_{t-1}^{(f)} + b_c) \tag{2.9}$$

$$h_t^{(f)} = o_t^{(f)} \cdot \sigma_h(c_t^{(f)}) \tag{2.10}$$

・Backward LSTM :

$$f_t^{(b)} = \sigma_g(W_f x_t + U_f h_{t+1}^{(b)} + b_f) \tag{2.11}$$

$$i_t^{(b)} = \sigma_g(W_i x_t + U_i h_{t+1}^{(b)} + b_i) \tag{2.12}$$

$$o_t^{(b)} = \sigma_g(W_o x_t + U_o h_{t+1}^{(b)} + b_o) \tag{2.13}$$

$$c_t^{(b)} = f_t^{(b)} \cdot c_{t+1}^{(b)} + i_t^{(b)} \cdot \sigma_c(W_c x_t + U_c h_{t+1}^{(b)} + b_c) \tag{2.14}$$

$$h_t^{(b)} = o_t^{(b)} \cdot \sigma_h(c_t^{(b)}) \tag{2.15}$$

そして，前方向 LSTM の隠れ状態ベクトル$h_t^{(f)}$と後方向 LSTM の隠れ状態ベクトル$h_t^{(b)}$を結合する．

$$h_t = \left[h_t^{(f)}; h_t^{(b)}\right] \tag{2.16}$$

２．４　問題点

従来手法のうち２.２節の手法では音声とテキストの融合により音声単体の場合と比較して幸福クラスや恐怖クラス，驚愕クラスの誤分類が大きく改善される．しかし，悲哀クラスと中立クラス，憤怒クラス，驚愕クラスとの混同が顕著である．一方で，２.３節の手法では，音声と表情の融合により音声単体の場合と比較して悲哀クラスや幸福クラス，憤怒クラスの誤分類が少ないという結果が得られている．

すなわち，従来の感情推定のアプローチでは，個々の手法において特定の感情の識別精度が低いという欠点がある．

２．５　むすび

本章では，二つの従来感情推定手法のモデルの構造（LSTM 分類器，Bi-LSTM 分類器）および，問題点について述べた．

第３章　　提案手法

３．１　　まえがき

本章では従来の LSTM 分類器を用いた音声・テキストによる感情推定手法に対し，新たに表情のモーダルを加えるとともに, より広範囲の時系列情報を捉えられる Bi-LSTM 分類器を採用したマルチモーダル感情推定モデルを提案する．

３．２　　提案手法

二つの従来手法を組み合わせることで, 分離が容易でなかったクラスの推定性能を改善し，後分類を削減できると考えられる．そこで提案手法では，２.２節のモデルに対し入力として新たに表情を追加し，２.３節の Bi-LSTM ネットワークを適用する．

提案手法のモデルは Bi-LSTM ネットワークとソフトマックス層から構成される．モデルの構造を図３.１に示す. 提案モデルでは音声とテキスト, 表情を連結したベクトルを Bi-LSTM ネットワークへ入力し，その出力をソフトマックス層に通すことで 6 感情（憤怒，幸福，悲哀，恐怖，驚愕，中立) のクラス分類を実行する．また，比較のため音声とテキスト，表情に加え，音声とテキスト，表情の融合のほかに音声のみ，テキストのみ，表情のみ，音声とテキスト，表情とテキスト，音声と表情の計 6 種類のベクトル組み合わせに関しても同様に連結し，ネットワークへの入力とした．

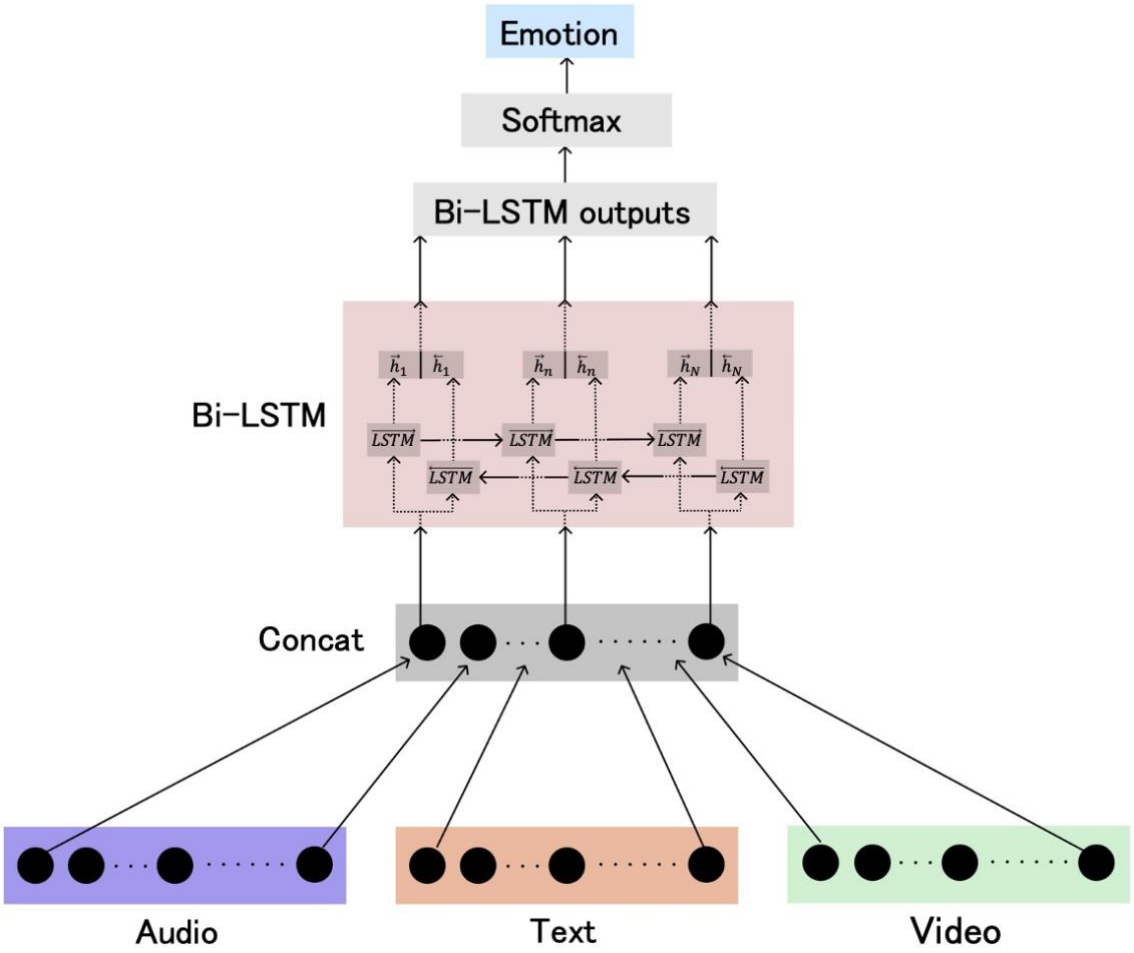

図　3.1　提案手法のモデル構造

3.3　むすび

本章では，本研究の提案手法である Bi-LSTM 分類器を用いた音声・テキスト・表情を入力とするマルチモーダル感情推定手法におけるモデルの構造や入力ベクトルの構成について述べた．従来の LSTM 分類器に対し双方向の時系列情報を付加するとともに，入力として新たに表情のモーダルを加えることで感情の誤分類を改善する．

第４章　実験

４．１　まえがき

本章では，本研究で用いるデータセット，実験方法，実験結果，考察について述べる．具体的には，分類器の学習に用いたデータセット，実験方法について述べ，評価実験としてそれぞれのモーダルを組み合わせたベクトルをモデルに入力した時の従来手法と提案手法の実験結果の比較を行う．

４．２　分類器の学習および精度比較

４．２．１　データセット

分類器の学習およびテストには Interactive emotional dyadic motion capture database (IEMOCAP) [9]を用いる．IEMOCAP は 10 人の被験者による五つのセッションを録画・録音した合計約 12 時間のオーディオ・ビジュアル情報である．音声だけでなく，テキスト（セリフ）や顔や頭，手などに取り付けられたマーカーの座標情報も含まれているが，本実験では表情（顔の各部位の 3 次元座標），音声，テキストの三つを用いる．また，本データセットはあらかじめ各セッションが平均 4.5 秒ほどの発話に分割されており，合計 10039 個の発話からなる．それぞれの発話に，幸福，興奮，悲哀，欲求不満，恐怖，驚愕，中立，嫌悪，その他の 9 種類の感情ラベルが振り分けられている．

４．２．２　データの前処理

(a) 表情

各発話の開始タイムスタンプと終了タイムスタンプに従い，55 部位 165 次元の座標の中央値を取得した．ただし，発話内の全ての時間において NAN を含む場合，その発話を除外した．また，先行研究と同様に感情クラスごとのサンプル数の不釣り合いを改善するべく，恐怖のクラスを 30 倍，驚きクラスを 10 倍にアップサンプリングした．さらに，興奮クラスと幸福クラスはその類似性より幸福クラスに統一した．本実験では用いない感情クラスを削除し，最終的には憤怒，幸福，悲哀，恐怖，驚愕，中立の 6 感情クラスからなる合計 8992 個の発話が含まれるデータセットを得た．各感情クラスのサンプル数を表 4 .1 に示す．最後に，80%をトレーニングセットとし，20%をテストセットとする．その他のモーダルと公平な比較ができるよう，全てのモーダルの実験で同様に分割する．

表 4.1 各感情クラスのサンプル数

Class	Count
Angry	1039
Happy	1502
Sad	2716
Fear	1085
Surprise	1089
Neutral	1561
Total	8992

(b) 音声，テキスト

各発話の開始タイムスタンプと終了タイムスタンプに従い, 以下に示す 8 次元の音響的特徴および 2339 次元の言語的特徴を取得した．それらの特徴に対し表情と同様の処理を施し，6 感情クラスからなる合計 8992 個の発話が含まれるデータセットを得る．そして，トレーニングセットとテストセットに分割する．

(1) ピッチ

ピッチとは音声信号の 1 秒あたりの振動回数に相当し, 人間の耳に音高として認識されるものである．例えば, 興奮しているときや怒りを感じているときにはピッチが上昇し，落ち着いているときや悲しんでいる場合には下降する．様々な推定アルゴリズムが存在するが, ここでは最も一般的であるセンタークリップされたフレームの自己相関に基づく方法[10]を使用する. センタークリップされた信号$y_{clipped}[n]$は式(4.1)で表される．

$$y_{clipped}[n] = \begin{cases} y[n] - C_l & (y[n] \geq C_l\ ) \\ 0 & (|y[n]| < C_l\ ) \\ y[n] + C_l & (y[n] \geq C_l\ ) \end{cases} \tag{4.1}$$

ここに，C_lは入力信号$y[n]$の平均のほぼ半分，$[n]$は信号の離散性である．ここで得られた$y_{clipped}[n]$に基づいて自己相関を計算し，さらに正規化を経てピーク値を得る．

(2) ハーモニクス

ハーモニクスとは怒りやストレスを感じている際などに現れるピッチ以外の励起信号であり，音の複雑さや音色を形成する追加的な周波数成分を指す．ここでは中央値フィルタを用いた計算法[11]を用いる．与えられたウィンドウサイズに基づいてスペクトログラムの周波数スライスS_hに式(4.2)で表される中央値フィルタを適用し，式(4.3)で表されるハーモニクスを強調したスペクトラムH_iを生成する．

$$y[n] = median(x[n-k:n+k]\ |k = (l-1)/2) \tag{4.2}$$

$$H_i = M(S_h, l_{harm}) \tag{4.3}$$

ここに，$x[n]$は入力信号，y$[n]$は出力信号，lは奇数のウィンドウサイズ，kはウィンドウの中心からの距離，l_{harm} はハーモニックフィルタの長さ，iはタイムステップである．

(3) 音声エネルギー

音声エネルギーとは音声信号の強度であり，ここでは標準的な Root Mean Square Energy (RMSE) を採用した. 例えば, 興奮した状態では大きな声で話す傾向があるため, 音声エネルギーの値は大きくなる．音声エネルギーEは信号の振幅$y[i]$を用いて式(4.4)で表される.

$$E = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y[i]^2} \tag{4.4}$$

(4) 休止

休止は音声信号内の静かな部分であり，閾値t ($\approx 0.4 * E$)を用いて式(4.5)で表される. 例えば，興奮した状態では速く話す傾向があるため，休止値は低くなる.

$$Pause = P_r\ (y[n] < t) \tag{4.5}$$

(5) 中心モーメント

中心モーメントとは信号の振幅の平均と標準偏差を用いて表される信号の全体的な特性を指す.

(6) Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF とは特定の単語やトークンが文書内でどのくらい重要であるかを示しており, Term Frequency (TF)と Inverse Document Frequency (IDF)から構成される．TF は文書内で特定の単語やトークンが出現する回数を示している．一方, IDF は冠詞など頻繁に出現する単語やトークンによるバイアスの影響をなくすためのものであり，式(4.6) で表される．TF-IDF 値は TF と IDF の積として計算される.

$$idf(t, D) = log\frac{N}{|d \in D : t \in d|} \tag{4.6}$$

ここに，tは特定の単語やトークン，Dは文書の集合，Nは文書集合内の文書数，dは特定のtを含む文書の集合とする.

(c) 複数のモーダルの組み合わせ

複数のモーダルから効率的にベクトルを融合する方法はいくつか提案されているが, 本実験では単純にそれぞれの特徴ベクトルを連結する. それぞれのモーダルのデータの発話の並びが同じになるよう連結することにより特徴ベクトルを取得し, トレーニングセットとテストセットに分割する.

４．２．３　　損失関数および学習パラメタ

学習時の損失関数には従来手法と同様にクロスエントロピー損失を用いる. 多クラス分類タスクにおけるクロスエントロピー損失とはモデルが出力した予測の確率分布と, 実際のラベルの分布との間の不一致を数値化したものであり, モデルのパフォーマンスを表す．クロスエントロピー損失の定義は式(4.7)で表される.

$$L(y,p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{i,c}\log\left(p_{i,c}\right) \tag{4.7}$$

ここに，N はサンプル数，M は感情クラスの数，$y_{i,c}$はサンプルiがクラスcに属する場合は 1，そうでない場合に 0 となるワンホットエンコーディングされたラベル，$p_{i,c}$はモデルによりサンプルiがクラスcに属すると予測される確率である．

この目的関数を最適化するために損失関数の値を最小化するようにパラメタを調整するアルゴリズムである Adam Optimizer を採用する．Adam Optimizer は損失関数から得られる勾配情報を利用して各イテレーションでモデルの重みを適切に更新する役割を持つ．

各実験で共通する各種学習パラメタは，LSTM および Bi-LSTM の層数 2,ドロップアウト[12]を 0.2，エポック数を 55000 である．ドロップアウトとは分類器の隠れ空間を正則化するためのシャットオフメカニズムであり，ネットワークのロバスト性を高める効果がある．その他のパラメタを表４.２に示す．

表 ４.２ 各モーダルにおけるパラメタ

Modal	Input dimension	Hidden dimension	Batch size	Learning Rate
audio	8	50	1567	10^{-3}
text	2339	500	128	10^{-4}
video	165	256	1567	10^{-4}
audio ＆ video	173	256	200	10^{-4}
video ＆ text	2504	256	200	10^{-5}
audio ＆ text	2347	256	200	10^{-3}
audio ＆ text ＆ video	2512	256	200	10^{-5}

４．２．４　　評価指標

本実験では評価指標として正解率（accuracy），適合率（Precision），再現率（Recall），F 値（F-measure）を用いた．多クラス分類タスクにおける予測値と真値の関係を表現した混同行列を図４.１に示す．

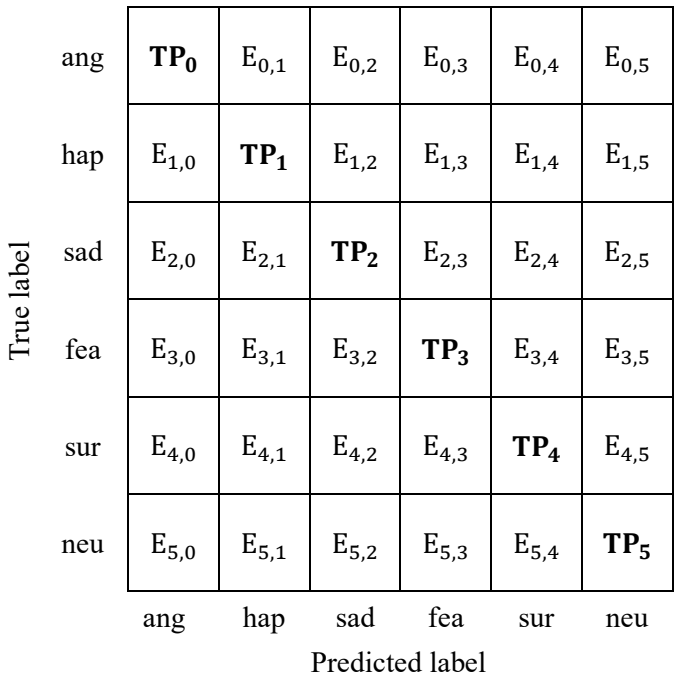

図 4.1 多クラス分類タスクの混同行列

TP_iはクラスiが正しく予想された数（True Positive），$E_{i,j}$はクラスiがクラスjに誤って予測された数（Error)である．

(a) 正解率（Accuracy）

全予測の中で正しく予測できた割合を示し，式(4.8)で表される．

$$Accuracy = \frac{\sum_{i=0}^{5} TP_i}{Total\ number\ of\ samples} \tag{4.8}$$

(b) 適合率（Precision）

クラスiと予測した中で実際にクラスiであった割合を示し，式(4.9) で表される．

$$Precision_i = \frac{TP_i}{TP_i + \sum_{j=0, j\neq i}^{5} E_{j,i}} \tag{4.9}$$

(c) 再現率（Recall）

実際にクラスiであるものの中でクラスiと予測できた割合を示し, 式(4.10) で表される．

$$Recall_i = \frac{TP_i}{TP_i + \sum_{j=0, j\neq i}^{5} E_{i,j}} \tag{4.10}$$

(d) F 値（F-measure）

適合率と再現率はトレードオフの関係にあるため，それらを一つにまとめるべく用いられる調和平均を示し，式(4.11) で表される．

$$F-measure_i = \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i} \tag{4.11}$$

４．２．５　評価結果

LSTM 分類器および Bi-LSTM 分類器において，各モーダルの組み合わせを入力としたときの感情分類の精度比較を行う．それぞれの分類器を用いた場合の評価結果を表４．３，表４．４に示す．また，各モーダルの混同行列を図４．２-図４．８に示す．従来手法を青，提案手法を赤でマーキングする．

表　４．３　各モーダルの評価結果(LSTM 分類器)

Modal	Accuracy	Precision	Recall	F-measure
audio	40.6	41.9	39.4	39.3
text	47.6	50.9	44.8	41.9
video	50.6	47.8	48.8	45.9
audio ＆ video	52.8	52.1	52.5	49.7
video ＆ text	66.9	66.9	69.2	66.2
audio ＆ text	**59.8**	**62.3**	**59.3**	**60.1**
audio ＆ text ＆ video	67.4	67.6	70.0	66.7

表　４．４　各モーダルの評価結果(Bi-LSTM 分類器)

Modal	Accuracy	Precision	Recall	F-measure
audio	48.9	51.8	46.2	46.8
text	63.1	64.4	64.6	63.9
video	55.5	55.3	52.1	51.4
audio ＆ video	53.7	53.3	53.9	52.0
video ＆ text	68.9	68.7	70.7	68.4
audio ＆ text	60.4	59.8	62.8	60.5
audio ＆ text ＆ video	**69.2**	**68.8**	**71.4**	**69.3**

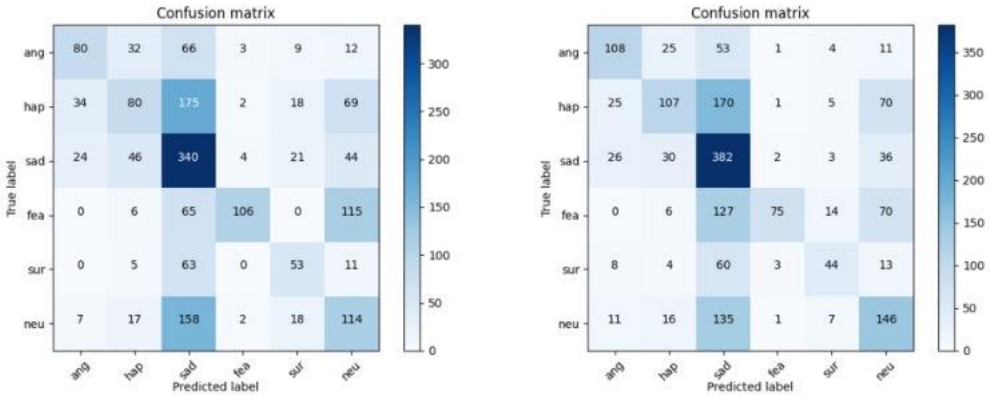

図 4.2 音声を入力としたときの混同行列（左:LSTM 分類器，右:Bi-LSTM 分類器）

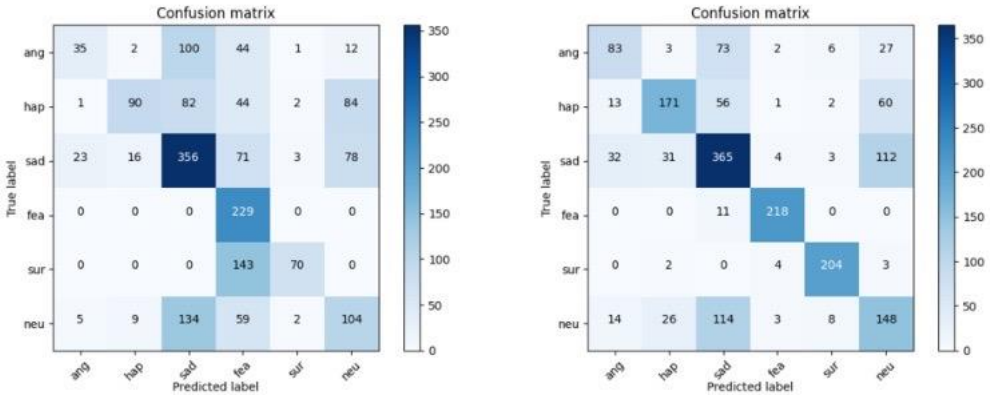

図 4.3 テキストを入力としたときの混同行列（左:LSTM 分類器，右:Bi-LSTM 分類器）

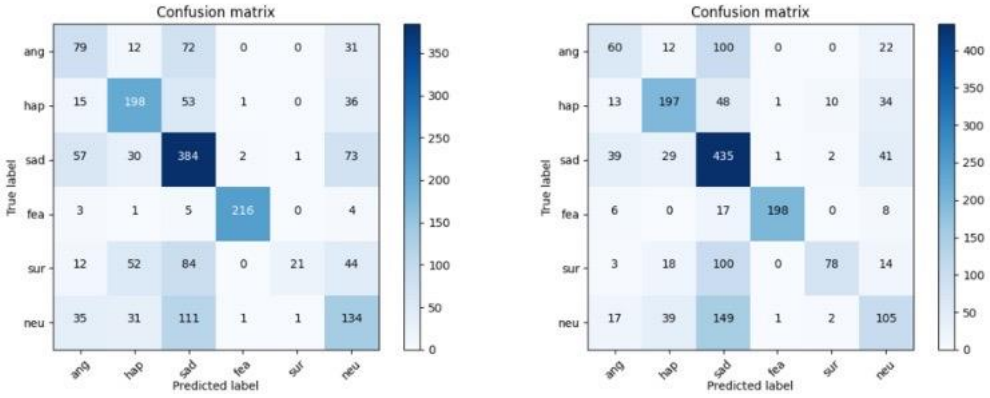

図 4.4 表情を入力としたときの混同行列（左:LSTM 分類器，右:Bi-LSTM 分類器）

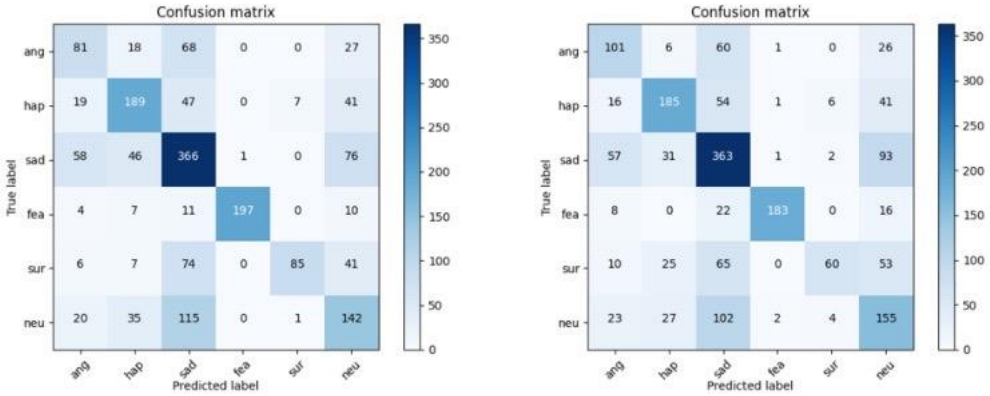

図 4.5 音声・表情を入力としたときの混同行列（左:LSTM 分類器，右:Bi-LSTM 分類器）

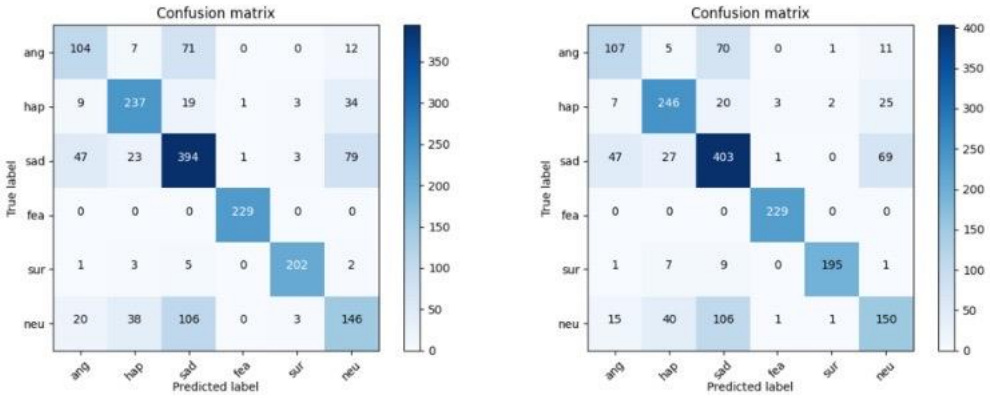

図 4.6 表情・テキストを入力としたときの混同行列（左:LSTM 分類器，右:Bi-LSTM 分類器）

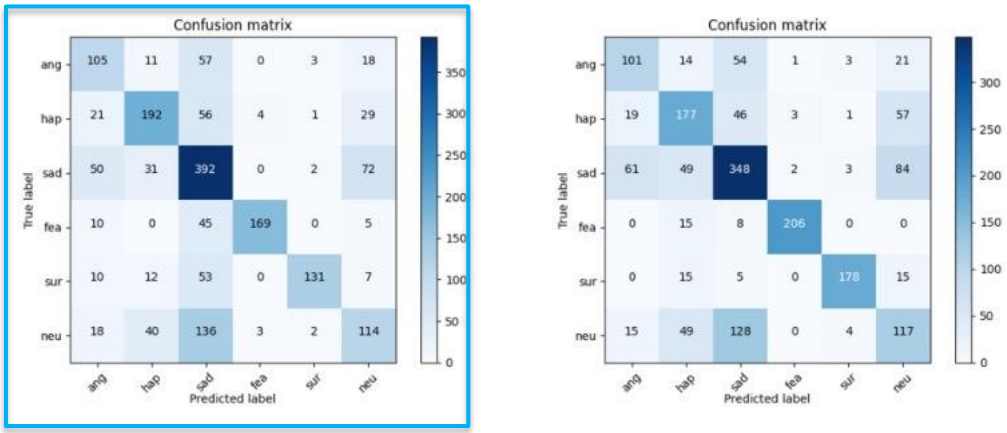

図 4.7 音声・テキストを入力としたときの混同行列（左:LSTM 分類器，右:Bi-LSTM 分類器）

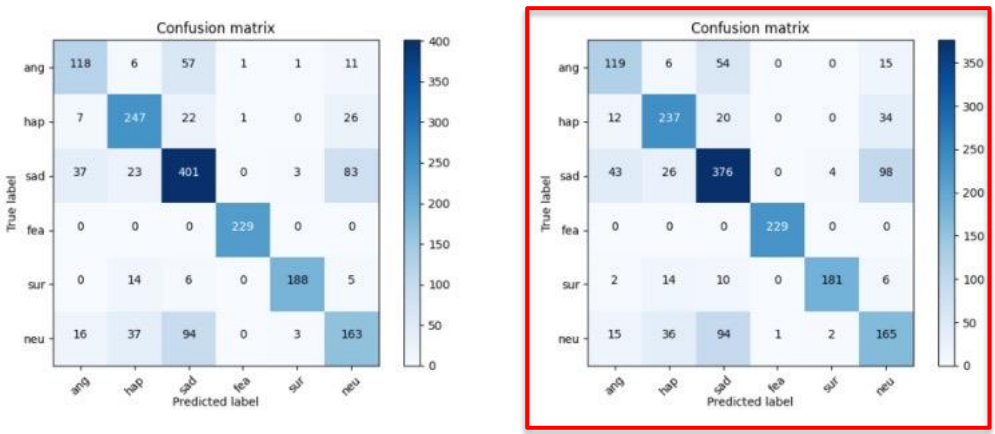

図 4.8 音声・テキスト・表情を入力としたときの混同行列（左:LSTM 分類器, 右:Bi-LSTM 分類器）

4.3 考察

まず，各モーダルの組み合わせ精度比較を行う．LSTM 分類器では，複数のモーダルを組み合わせることでおおよそ各指標が上昇しており，とりわけ音声と表情および音声とテキストの組み合わせで飛躍的に上昇した．このことより，テキストの寄与による恩恵が極めて大きいことがわかる．一方で Bi-LSTM 分類器では，各モーダルに音声を加えると下がってしまう場合がほとんどであるため，モーダルの結合方法や用いる音響的特徴を見直す必要があると考えられる．

次に，モーダルの組み合わせごとに LSTM 分類器と Bi-LSTM 分類器の精度比較を行う．単一モーダルについてはいずれも各指標が上昇したが，とりわけ大幅な上昇が見られたのがテキストであった．これは過去から未来への情報処理だけでなく，未来から過去への情報処理も可能になったことで，文脈をより的確に把握できたことに起因すると考えられる．その他の複数のモーダルの組み合わせではわずかな上昇しか見られず，混同行列についても大きな変化は見られなかった．各特徴ベクトルの融合方法や NAN のみを含む発話の削除により未来の時系列情報の恩恵を受けることができなかったことが原因であると考えられる．

さらに，従来の LSTM 分類器を用いた音声・テキストによる感情推定手法と提案手法である Bi-LSTM 分類器による音声・テキスト・表情による感情推定手法の精度比較を行う．正解率が 9.4%，適合率が 6.5%，再現率が 12.1%，F 値が 9.2%上昇したが，より変化が顕著に見られたのは混同行列であった．従来手法で問題であった悲哀クラスと憤怒クラス，驚愕クラスの混同は大きく改善されたが，悲哀と中立の混同に関してはあまり改善が見られなかった．また，悲哀以外の全ての感情で正しい感情を予測できる割合が上昇したが，とりわけ恐怖の上昇が大きい．これは恐怖という感情が表情の特徴として現れやすいことに起因していると考えられる．以上より提案手法では各感情の誤分類が改善したことが確認できた．

４．４　むすび

本章では，本研究の感情推定モデルにおける，分類器の学習に用いたデータセット，実験方法について述べた．また，従来手法と提案手法の実験結果の比較を行った．提案手法では感情推定の精度の改善が達成でき，有効性を示すことができた．

第５章　結論と今後の課題

５．１　結論

本研究では，従来の LSTM 分類器を用いた音声・テキストによる感情推定手法に対し，双方向の時系列情報を扱える Bi-LSTM 分類器に変更しかつ，入力に表情を加えたマルチモーダル感情推定手法を提案した. 提案手法では双方向の時系列情報や視覚的特徴を加味することで 6 感情をより正確に分類することが可能となった. 実験では各モーダルの組み合わせおよび各分類器を用いた場合の精度を比較することにより, 提案手法の有効性を確かめた.

５．２　今後の課題

提案手法では表情を入力に追加することによる大きな精度改善を確認できたが双方向の時系列情報の付与による精度改善は比較的小さい. そのため，今回は行えなかったモーダルの融合方法の検討が不可欠である. また，音声を入力に追加することによる精度低下が一部見られたことから今回用いた音響的特徴に加え，Mel-Frequency Cepstral Coefficients (MFCC)[13]などの周波数領域の特徴を含めることにより音声モーダルの追加による精度改善が見込めると考えられる.

さらに, 本研究で用いた表情のデータセットは音声やテキストとベクトルの大きさを揃えるために各発話の 55 部位 165 次元の座標の中央値としたため，時系列情報を上手く加味することができなかった. そのため，双方向の時系列情報を最大限に活かせるような表情データの処理方法を模索する必要がある.

謝辞

本論文の執筆に当たり，研究の方向性や問題点をご指導くださり，快適な研究環境を与えてくださった渡辺裕教授に感謝いたします.

また，日頃から興味深い研究内容を共有していただきかつ，丁寧なアドバイスをくださった渡辺研究室の皆様に感謝いたします.

最後に，私をここまで育ててくださり，常に心を支えてくださり，生活を支えてくださっている家族に感謝いたします.

參考文獻

[1] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, R.W. Picard, “Driver Emotion Recognition for Intelligent Vehicles,” ACM Computing Survey, vol. 53, pp. 1-30, Jul. 2020.

[2] A.C. Nyquist, A.M. Luebbe, “An Emotion Recognition-Awareness Vulnerability Hyp othesis for Depression in Adolescence: A Systematic Review,” Clinical Child and F amily Psychology Review, vol. 23, pp. 27-53, Mar. 2019.

[3] C. Greco, O. Matarazzo, G. Cordasco, A. Vinciarelli, Z. Callejas, A. Esposito, “Dis criminative Power of EEG-Based Biomarkers in Major Depressive Disorder: A Syst ematic Review,” in IEEE Access, vol. 9, pp. 112850-112870, Sep. 2021.

[4] S. Argaud, M. Vérin, P. Sauleau, D.Grandjean, “Facial emotion recognition in Parkinson's disease: A review and new hypotheses,” Movements Disorders, vol. 33, pp. 554-567, Feb. 2018.

[5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[6] Gaurav. Sahu, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," CoRR, abs/1904.06022, Apr. 2019.

[7] S.E. Romero, R. Kleinlein, C.L. Jiménez, J.M. Montero, F.F. Martínez, “GTH-UPM at DETOXIS-IberLEF 2021: Automatic Detection of Toxic Comments in Social Networks,” In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), Co-Located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), Málaga, Spain, vol. 2943, pp. 533–546, Sep. 2021.

[8] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J.M. Montero, F. Fernández-Martínez, "A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset," Applied Sciences, vol. 12, no. 1, p. 327, Dec. 2022.

[9] C. Busso, M. Bulut, C-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, Nov. 2008.

[10] M. Sondh, "New methods of pitch extraction," IEEE Transactions on Audio and Electroacoustics, vol. 16, no. 2, pp. 262-266, Jun. 1968.

[11] Derry. FitzGerald, ”Harmonics/Percussive Separation Using Median Filtering,” Proceedings of the International Conference on Digital Audio Effects (DAFx), Graz, Austria, pp. 246-253, Jan. 2010.

[12] Leo. Breiman, “Bagging predictors,” Machine Learning, vol. 24, no. 2, pp. 123-140, Aug. 1996.

[13] S. Davis, P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech,

and Signal Processing, vol. 28, no. 4, pp. 357-366, Aug. 1980.

図一覧

表一覧