マルチモーダル深層学習を用いた発話内容の信頼性推定

木戸　喜隆† 　山田　一権†† 　白井　匡人†††

† 島根大学 自然科学研究科　〒 690-8504 島根県松江市西川津町 1060
†† エクスウェア株式会社　〒 140-0002 東京都品川区東品川 4 丁目 10
††† 島根大学 学術研究院理工学系　〒 690-8504 島根県松江市西川津町 1060
E-mail: †n19m103@matsu.shimane-u.ac.jp, ††k-yamada@xware.co.jp, †††shirai@cis.sihmane-u.ac.jp

あらまし　本研究では, 被験者の発言内容の嘘を検出することを目的とし, マルチモーダル深層学習を用いて被験者の発話内容の信頼性を推定する. 近年, 嘘検出に関する研究では, 単一のモーダルではなく, 映像, 音声, テキストなどの複数のモダリティを用いることで精度が向上することが報告されている. しかしながら, 嘘検出に関する研究では十分な量のデータセットが公開されておらず, 多くの研究ではその効果は限定的にしか検証されていない. 評価実験では, 各被験者に対して真実と嘘を区別して発言したデータセットを作成し, 発話内容の信頼性の評価を基に提案手法の有効性を検討する.

キーワード　マルチモーダル深層学習，信頼性予測

1 前　書　き

人が面接や裁判といった場面で意図的に嘘をつくという行為は社会に大きな影響を与える可能性がある. したがって, 重要な場面で嘘を正確に検出することは重要である. しかしながら, 人の嘘を検出することは困難であり, BondJr [1] らの研究によると, 専門家でない場合, 54%の精度でしか嘘を検出できないことが示されている.

また近年, AI の急速な発展に伴い, 動画から嘘を検出するという試みが増えている [2] [3]. 先行研究の多くは, 裁判で得られた映像, 音声, テキストのマルチモーダルなデータセット [4] を使用している. しかし, 先行研究は交差検証中に精度のスコア付けを調整, また, 各モダリティで学習するデータにテストデータを含むなど誤った方法で実験を行っている. 実際に, Rill-Garcia らの研究 [5] では先行研究と同様のデータセットを用いて各モダリティの精度を検証しているが, 先行研究と比較して精度が大きく劣っている. したがって, 嘘が検出できるか明らかになっていない.

そこで本研究では, 映像, テキスト, 音声の 3 つのモダリティから特徴を抽出して発話内容の信頼性を評価できるか検討する. そのため, 提案手法として, 映像, テキスト, 音声それぞれ各モダリティにおいて LSTM モデルを用いて特徴量を抽出し, それらを全結合してマルチモーダル深層学習を行う方法を提案する. 提案手法を用いて学習データに対するテストデータの正解率を評価する.

第 2 章では, 嘘検出に関する先行研究について述べ, 第 3 章ではマルチモーダル深層学習について述べる. 第 4 章では提案手法について述べる. 第 5 章では実験により提案手法の有効性を示し, 第 6 章で結論を述べる.

2 嘘　検　出

2.1 関 連 研 究

嘘検出をトピックとした研究は近年, 非常に盛んである.

Wu らの研究 [6] は, 嘘を検出する為に, マルチモーダル深層学習を行う手法を提案している. 彼らの実験では, 121 本の裁判のデータセットを用いて映像, テキスト, 音声, 顔の微細な動き（Micro-Expression）の 4 つの特徴量を抽出し, 嘘検出を行っている. Krishnamurthy らの研究 [3] は, Wu らの研究を基に, ニューラルネットワークを用いて嘘検出を行っている. 彼らの研究は, Wu らと同様のデータセットに対し, それぞれ映像, テキスト, 音声モダリティでで 3D-CNN, TextCNN, CNN を使用して嘘検出を行っている.

しかし, Wu らの研究では交差検証中で分類結果が最も良くなるように各モダリティのスコア付けを調整しており, Krishnamurthy らの研究でも各モダリティで学習する際にテストデータも学習に使用するという誤った方法で実験を行っている. したがって, 嘘検出の研究において, 嘘が正しく検出できるのかが明らかになっていない.

2.2 問 題 設 定

Ding らの研究 [7] で人が嘘をつく際に顔と体の動きは非同期的であると示されている. しかし, Wu らや Krishnamurthy らの研究では, 時系列を考慮した学習を行っていない. そこで, 本研究では時系列を考慮する学習モデル LSTM を用いて発話内容の信頼性推定を行えるか明らかにする.

3 マルチモーダル深層学習

マルチモーダル深層学習は, 映像, テキスト, 音声などの複数のモダリティを統合的に処理するので, 複数の要因を考慮した予測をすることが可能となる手法である. 以下に, Wu らと

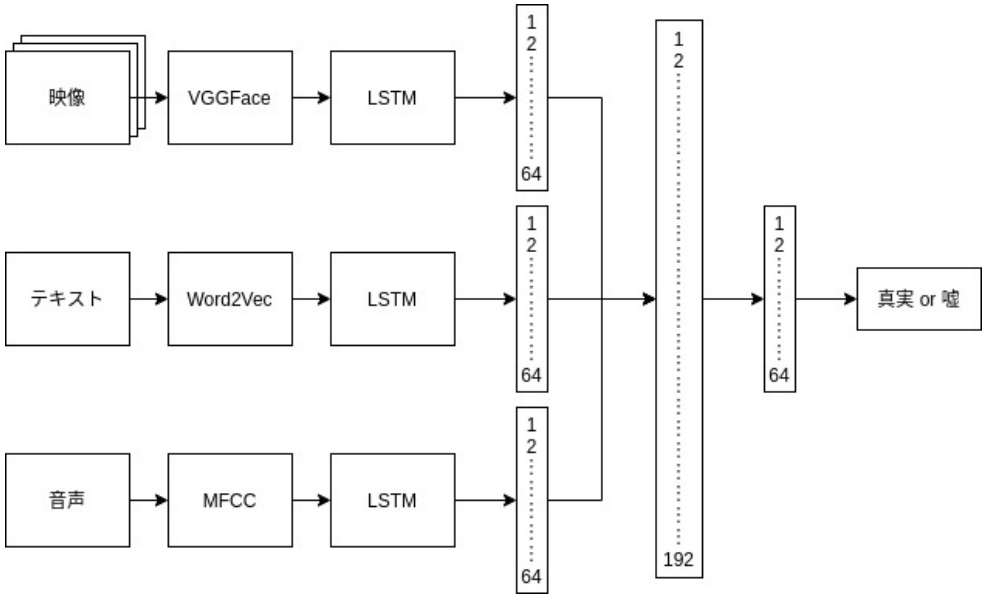

図 1　モデル概要

Krishnamurthy らがそれぞれのモダリティで使用している特徴量抽出の手法を示す.

3.1 映　　像

Wu らの研究では, Improved Dense Trajectory（IDT）[8] を用いて特徴量抽出を行っている. また, Krishnamurthy らの研究では, 3D-CNN を用いて 300 次元のベクトルを抽出を行っている.

3.2 テキスト

Wu らの研究では, Glove [9] を用いて単語を学習し, 300 次元のベクトルへ変換を行っている. また, Krishnamurthy らの研究では, TextCNN を用いて 300 次元のベクトル抽出を行っている.

3.3 音　　声

Wu らの研究では, 音声データを MFCC 特徴量に変換し, Krishnamurthy らの研究では, OpenSMILE [10] を用いて特徴量を抽出し, 多層パーセプトロンを用いて 300 次元のベクトルへと変換する.

3.4 損失関数

学習時にモデルの出力と正解ラベルのクロスエントロピーを最小限に抑える必要がある. そこで損失関数を以下のように, Krishnamurthy らは定めている.

$$J = \frac{-1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log_2(\hat{y}_{ij}) \qquad (1)$$

ここで, N はサンプル数, C はカテゴリの数を表す. また, y_{ij} は i 番目の One-Hot ベクトルの正解ラベルであり, $\hat{y}_{ij}$ はクラス j に属すると予測される確率である.

4 提案手法

提案手法は, 映像, テキスト, 音声それぞれのモダリティで特徴量を抽出し, 時系列を考慮する学習モデル Long Short Term Memory（LSTM）を用いてマルチモーダル深層学習を行い, 発話内容の信頼性推定を行うというものである. 本実験で設定している各モダリティの特徴量抽出の手法を以下に示す.

4.1 映　　像

本研究では, 1 本の動画から 5 枚の映像（フレーム）を等間隔に抽出する. 抽出した 5 枚のフレームは高さ 112, 幅 112 のカラー画像（112, 112, 3）であり, 学習済み VGG-Face CNN [11] モデルを用いて 512 次元のベクトルに変換する. それらを LSTM モデルを通して, 64 次元のベクトルに変換する.

4.2 テキスト

テキストは動画に Google Cloud Speech API を用いて文字起こしを行ったデータを使用する. 抽出したテキストを Wikipedia の日本語学習済みモデルを用いて単語ごとに 200 次元のベクトルに変換する. それらを LSTM モデルを通して, 64 次元のベクトルに変換する.

4.3 音　　声

音声は全ての映像に対し, Mel-frequency Cepstral Coefficients（MFCC）特徴量を抽出する. 抽出した MFCC 特徴量を LSTM を用いて 64 次元のベクトルに変換する.

4.4 結　合　層

結合層では, 映像, テキスト, 音声それぞれ 64 次元の特徴量を全結合層の入力とし, 真実か嘘かの 1 次元を出力する. 出力

層の活性化関数にはシグモイド関数を適用する．本提案手法では，全ての LSTM モデルにおいて，過学習を防ぐためにドロップアウト（$p = 0.15$）を設定する．

5 評価実験

実験では，映像，テキスト，音声それぞれのモダリティで特徴量を抽出し，LSTM を用いてマルチモーダル深層学習を行う．提案手法の有効性を確認するため，本研究で作成したデータセットを用いて，先行研究の手法と本研究の手法での正解率を比較する．

5.1 データセット

本研究は面接時における嘘を検出することを想定してデータセットの作成を行った．したがって，データセットには被験者である 54 名の大学生が模擬面接を受けている映像，テキスト，音声データが含まれる．模擬面接は表 1 の手順で行う．

表 1　模擬面接の流れ

面接官	被験者
（面接の概要を説明）	
以上のことを理解していただけましたか？	
	はい
（録画開始）	
それでは模擬面接を開始します．	
（質問 1~6）	
	（回答 1~6）
以上で前半は終了です．	
これから後半に移ります．	
「嘘」や「話を盛らず」に本心でお答えください．	
	はい
それでは後半を始めます．	
（質問 1~6）	
	（回答 1~6）
（録画終了）	
以上で模擬面接を終わります．	

実験環境は図 2 のようになっており，面接者は被験者に向かって正面に座る．被験者とカメラ（スマートフォン）との距離は 100cm であり，被験者の声を鮮明に録音するため，被験者にはピンマイクを装着している．面接官は被験者に予め，6 問の質問を行うこと，前半と後半で同じ質問を繰り返すこと，前半には嘘や話を盛って回答し，後半は嘘や話を盛ること無く本心で回答することを伝える．被験者の了承が得られたら，面接官はカメラを起動し録画を開始する．面接官は表 4 に示す 6 問を被験者に質問する．前半と後半の両方を終えたら，面接官はカメラの録画を終了する．模擬面接の後，被験者にアンケート調査を行い，前半のパートでどの程度を嘘をついたのか 5 段階で評価してもらう．

データセットは 1 人の学生に対し，真実と嘘を区別して回答

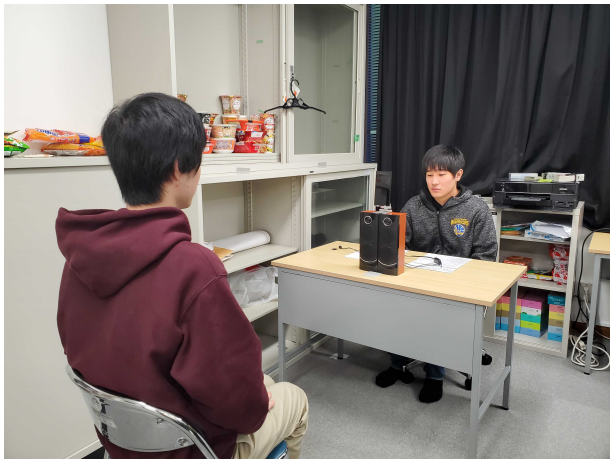

図 2　実験環境

した 12 本の動画データが含まれており，データセットの総数は 647 本の動画と 54 名分のアンケート結果が含まれる．

5.2 実験設定

比較手法にはニューラルネットワークを用いてマルチモーダル深層学習を行う手法 [3] を用いる．本実験で使用するデータセットは 647 本の動画の内，被験者 54 名がアンケートで最も嘘を盛ったと回答した 114 本の嘘の動画，かつ，324 本の真実の動画合わせて 468 本を使用する．本実験に使用した全モデル，各モダリティでのパラメータを表 3 に示す．

5.3 評価方法

実験の評価には正解率（Acc）を用いる．実験では学習データとテストデータを分割する際，人による重複は避けなければならない．したがって，実験では被験者による 9 分割交差検証を行う．これにより，同じ被験者が学習データとテストデータに含まれるという可能性を無くすことできる．

5.4 実験結果

表 2 に実験結果を示す．映像特徴量でのみの正解率は 62.1%だった．また，テキスト特徴量のみの正解率は 64.5%であり，音声特徴量のみの正解率は 65.5%であった．マルチモーダル学習を行った際の正解率は 65.3%であった．

表 2　実験結果

Features	Acc	
	比較手法	提案手法
映像	65.5	69.1
テキスト	64.4	64.5
音声	62.4	65.5
All Features	66.7	68.7

表 3　使用したパラメータ

Features		全モデル	各モダリティ
活性化関数	中間層	ReLU	ReLU
	出力層	sigmoid	sigmoid
損失関数		二値クロスエントロピー	二値クロスエントロピー
最適化関数		Adam	Adam
バッチサイズ		32	32
エポック数		50	50

表 4　質 問 項 目

質問 1	あなたの長所を教えてください.
質問 2	大学の授業以外で自主的に勉強していることを教えてください.
質問 3	綺語を選ぶ上で最も大事なのは次の内どれですか, 社風, 給与, 仕事内容, 自己成長.
質問 4	入社して配属された先の仕事が, あなたが最もやりたくない仕事でした. あなたはどうしますか.
質問 5	一ヶ月以内のうちに行った, 良い行いを 1 つ教えてください.
質問 6	採用試験でグループワークを行うことになりました. あなたはどのような立場でグループワークに関わりますか.

5.5 考　　察

表 2 より, 提案手法を用いて映像, テキスト, 音声の各モダリティでのみ嘘を検出した場合, 映像モダリティを扱った場合が最も Acc が高いことがわかる. これは, 見かけの情報が最も人の嘘を検出する際に有効であることを示している. また, 全てのモダリティにおいて比較手法より提案手法の精度が向上していることがわかる. 更に, 全てのモダリティを全結合してマルチモーダル深層学習を行った場合, 比較手法と比べて提案手法の精度が 2%向上している. これは, 提案手法の有効的であることを示している.

6 結　　論

本研究では, 被験者の発言内容の嘘を検出するために, マルチモーダル深層学習を用いて被験者の発話内容の信頼性を推定する手法を論じた. 提案手法を用いて嘘検出を行った結果, 提案手法は, 既存研究と比較して Acc が 2%改善する. これにより, 提案手法の有効性を示した.

文　　献

[1] C.F. Bond Jr and B.M. DePaulo, “Accuracy of deception judgments,” Personality and social psychology Review, vol.10, no.3, pp.214–234, 2006.

[2] C. Bai, M. Bolonkin, J. Burgoon, C. Chen, N. Dunbar, B. Singh, V.S. Subrahmanian, and Z. Wu, “Automatic long-term deception detection in group interaction videos,” arXiv preprint arXiv:1905.08617, 2019.

[3] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, “A deep learning approach for multimodal deception detection,” arXiv preprint arXiv:1803.00344, 2018.

[4] V. Perez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, “Deception detection using real-life trial data,” In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp.58–66, 2015.

[5] R. Rill-Garcia, H. Jair Escalante, L. Villasenor-Pineda, and V. Reyes-Meza, “High-level features for multimodal deception detection in videos,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.0–0, 2019.

[6] Z. Wu, B. Singh, L.S. Davis, and V. Subrahmanian, “Deception detection in videos,” 2018.

[7] M. Ding, A. Zhao, Z. Lu, T. Xiang, and J.-R. Wen, “Face-focused cross-stream network for deception detection in videos,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.7802–7811, 2019.

[8] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, “A robust and efficient video representation for action recognition,” International Journal of Computer Vision, vol.119, no.3, pp.219–238, 2016.

[9] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp.1532–1543, 2014.

[10] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” Proceedings of the 21st ACM international conference on MultimediaACM, pp.835–838 2013.

[11] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al., “Deep face recognition.,” bmvc, vol.1, p.6, 2015.