感情情報で制御する FiLM ベース対話破綻検出モデル

中畔彪雅 [1,2]　吉野幸一郎 [3,2,1]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所
[3] 東京科学大学
nakaguro.hyuga.nl1@is.naist.jp　koichiro@c.titech.ac.jp

概要

音声対話では，同一の発話でも含まれるパラ言語情報によりニュアンスが変化し，不適切な応答は対話破綻を招く．本研究は，パラ言語情報から表出する感情に起因する感情的対話破綻の検出を目的とする．五つの感情を付与した発話と対応応答からなる paraling-dial を用い，発話感情と応答内容の不整合を判定するタスクを設定した．感情をテキスト意味を変調する制御信号とみなし，Feature-wise Linear Modulation (FiLM) を導入した結果，精度 89.5%を達成した．さらに，制御信号としては抽象度の高い感情ラベルが最も有効であることが示された．本研究は，感情的対話破綻検出における FiLM の有用性と，制御信号設計における抽象度の重要性を示す．

1　はじめに

ChatGPT に代表される大規模言語モデルの発展により，人間と自然に対話できる対話システムが実現されつつある [1]．これらのシステムはテキスト応答生成では高性能である一方，発話音声に含まれるトーンやニュアンスなどのパラ言語情報を十分に考慮できていない [2, 3]．そのためのモデルとして，音声を直接入力とする Speech Language Model (SLM) [4] や，音声特徴を抽出する Encoder [5] が提案されている．しかし，これらのモデルが実際にパラ言語情報をどの程度活用できているかを評価するためのベンチマークは十分に整備されていない．

対話の一貫性を評価する指標として，対話破綻検出は重要なベンチマークタスクである [6]．従来研究では，主にテキスト内容に基づく対話破綻が対象とされてきた．これに対し本研究では，パラ言語情報から表出する感情の不一致に起因する対話破綻を**感情的対話破綻**と定義し，パラ言語レベルでの対話破綻検出タスクを新たに定義する．これは，応答テキストが論理的に適切であっても，感情の乖離により対話品質が低下する現象を指す．

本研究では，パラ言語情報に由来する対話のニュアンスを捉えるためのデータセット **paraling-dial** を構築した．同一の発話内容に対して異なる感情を付与した音声データを用い，さらに感情ラベルと応答文をランダムにシャッフルすることで，感情的対話破綻を含むベンチマークデータを作成した．このデータを用いて既存の SLM [4] による対話破綻検出実験を行ったが，性能はチャンスレベルにとどまった．この結果は，既存の SLM では感情的対話破綻を検出する能力が十分に獲得されておらず，パラ言語情報とテキスト情報の統合方法にも課題があることを示唆している．

そこで本研究では，感情ラベルがテキスト解釈を変調させる制御信号として機能するという仮定を置き，Feature-wise Linear Modulation (FiLM) [7] を用いた感情的対話破綻検出器を構築した．実験の結果，提案手法が感情的対話破綻検出タスクに有効であることを確認し，今後の SLM 設計および学習に対する示唆を得た．

2　関連研究

2.1　テキストを対象とした対話評価

BLEU [8] や ROUGE [9] などの参照応答ベース指標は人手評価との相関が必ずしも高くなく，その代替として対話文脈への適合性に着目した対話破綻検出が提案されてきた．

対話破綻研究では，発話内容の細部よりも，応答全体が対話コンテキストに適合しているかという二値的評価が重視されてきた．

2.2　音声・パラ言語情報を用いた評価

HuBERT [5] などの音声事前学習モデルは，多様な音声タスクで高い性能を示している．

音声発話が持つこうしたパラ言語情報によって

表現されるニュアンスは，従来，感情ラベルに依拠した研究が多い．近年では，WavReward [10] のように，知的側面と情緒的側面を包括的に考慮しながら対話品質を評価する試みが見られる．

本研究では，これまでテキストで行われてきた対話破綻検出のアイデアを用い，特に発話が持つ感情的なニュアンスと，それに対応する応答とのミスマッチを検出可能かを評価する感情的対話破綻検出タスクを提案する．

2.3 マルチモーダル情報統合の課題

マルチモーダル情報の統合利用に関しては Early Fusion や Late Fusion などの手法が存在する．ただし，いずれも感情的対話破綻の検出には必ずしも適していない [11].

- Early Fusion: モダリティ間の情報量不均衡が生じやすい
- Late Fusion: 各モダリティを独立に処理するため，繊細な特徴を失いやすい

これらに対し，本研究で採用する Feature-wise Linear Modulation (FiLM) アーキテクチャ [7] は，音声由来のパラ言語情報を制御信号として用い，テキスト表現の中間層にスケーリングとシフトを動的に適用する．この構造により，感情的ニュアンスがテキスト解釈へ直接的に影響を与える統合が実現される．すなわち，FiLM は従来の特徴の結合ではなく，テキスト意味の変調による統合を可能にする点で異なる．

2.4 感情データセット

既存の感情音声対話データセット，例えば IEMOCAP [12] や MELD [13] では，感情の変化に伴って発話テキストも連動して変化する傾向がある．そのため，パラ言語情報のみが応答選択に与える影響を独立して評価することは困難である．パラ言語情報の影響を厳密に分析するためには，同一のテキスト内容に対して音声のニュアンス，例えば感情ラベルなどが変化し，その変化に応じて応答の適切性が異なるようなデータセットが不可欠である．このような条件を満たすデータセットの整備は，感情的対話破綻検出タスクの基盤として重要である．

3 感情的対話破綻検出タスク

発話のテキスト内容は同一であっても，発話が持つ音声ニュアンスにより発話—応答のペアが破綻しうる場合を想定する．

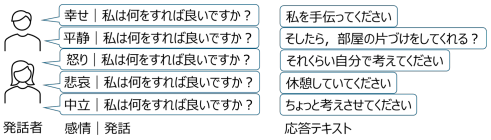

図 1 paraling-dial データ構造

3.1 paraling-dial データセット

本研究で構築した paraling-dail データセットの基本構造を図 1 に示す．本データセットは，単一のテキスト発話文に対して，話者が 5 種類の感情（幸せ・平静・怒り・悲しみ・中立）を込めて発話し，各音声に対して人手で応答文を付与したものである．従って，同一のテキストに対して感情の数だけ，発話音声と感情に応じた異なる応答文のペアが対応付けられている．データセットの構築手順は，以下の 4 段階で実施した．

1. 発話文の収集　著作権フリーで多様な表現がある青空文庫から対話表現である短文を 149 件収集し，ユーザの発話文とした．

2. 感情設定　ラッセルの感情円環モデル [14] に基づき，代表的な 5 感情（幸せ・平静・怒り・悲しみ・中立）を設定した．

3. 感情を考慮した応答文の作成　感情の変化に応じた各発話文に対する応答文を収集した．これらの応答文は，訓練された 1 名の作業者によって作成された．応答文は，ChatGPT による候補生成を参考にしつつ，訓練された作業者が感情差を反映するよう作成した．

4. 発話の収録　感情の差異を強調して表現できる声優または演劇経験者の男性 3 名と女性 3 名の話者により，各発話文へ感情を付与した発話を収録した．録音は防音室環境で指向性マイク[1)]を使用した．話者には，図 2 を示し，この図中の位置関係を基に各感情間の差異を表現するように指示した．最終的に，発話文 149 件× 5 感情× 6 話者の組み合わせにより，合計 4,470 組の発話と応答文のペアデータが収集された．総収録時間は約 284.89 分，1 発話あたりの平均時間は約 3.92 秒であった．

1） Sony ECM-674 エレクトレットコンデンサーマイク

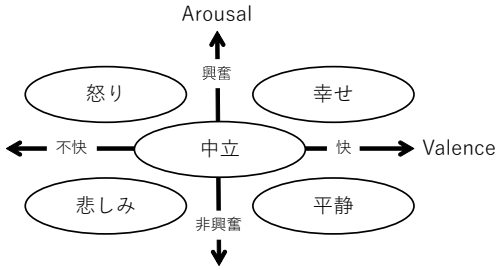

図 2　被験者に示した感情相関図

3.2　音響的妥当性の分析

実際に paraling-dial に含まれる音声が，意図した感情を音響的に適切に反映しているかを確認するため，各感情ラベルにおける基本周波数（F0）と実効値エネルギー（RMS エネルギー）の分布を分析した．感情ごとに明確な特徴的傾向が確認され，paraling-dial はモデル評価に十分な音響的妥当性を備えていることが確認された．

3.3　感情的破綻データセットの構築

paraling-dial は，同一発話文に対して複数の感情ラベル付き音声と，それぞれに対応する応答文から構成される．応答文を異なる感情に対応するものと入れ替えることで，疑似的な感情的対話破綻例を作成する．対話成立，破綻の組み合わせは以下の通り．

- 成立：特定の感情が込められた発話音声と，その感情に対応する正解応答文のペア
- 破綻：特定の感情が込められた発話音声と，異なる感情に対応する応答文を意図的に組み合わせたペア

これにより，テキストを固定したまま感情のみを操作し，モデルの感情的対話破綻への対処能力を評価できる．

4　感情によるテキスト解釈変調

本研究で定義した感情的対話破綻の検出には，発話のパラ言語情報が持つニュアンスやトーンを適切に取り扱う手法が必要である．既存の SLM では，こうした情報が必ずしも適切に活用されていない．そこで本研究では，感情がテキストの解釈を動的に変調する制御信号であるという仮説に基づき，Feature-wise Linear Modulation (FiLM) を用いた感情的対話破綻検出モデルを新たに提案する．

4.1　FiLM による変調モデル

本研究では感情を制御信号としてテキスト解釈を変調するという仮説を具体化するため，Feature-wise Linear Modulation (FiLM) を用いた．本モデルは，テキスト特徴抽出器，FiLM レイヤー，分類ヘッドの 3 つの主要コンポーネントから構成される．

まず，発話がもつテキスト情報を抽出するため，入力発話テキストと応答候補を「[SEP]」トークンで連結し，事前学習済み Sentence-BERT [15] に入力する．これにより，768 次元の特徴ベクトル x を得る．次に，制御信号として，感情ラベル（幸せ・平静・怒り・悲しみ・中立）を表す 5 次元の one-hot ベクトル c を用いる．これはテキストに付随するパラ言語的情報を高レベルに抽象化した表現である．これ以外にも，様々な抽象度で抽出したパラ言語情報に関連する特徴量を比較評価する．評価の詳細は 5 章で述べる．FiLM レイヤーでは，制御信号 c を全結合層に入力し，スケールパラメータ γ とシフトパラメータ β を動的に生成する．これらを用いてテキストベクトル x を以下のアフィン変化により変調する．

$$FiLM(x, c) = \gamma(c) \odot x + \beta(c) \tag{1}$$

この処理により，テキスト特徴 x は感情情報 c に依存して表現空間が適応的に変化する．最後に，分類ヘッドでは，変調後の 768 次元ベクトルを入力として，Linear(768 → 128) → ReLU → Dropout(0.3) → Linear(128 → 1) → Sigmoid からなるネットワークにより処理し，対話が破綻しているか否かの確率を出力する．

5　実験設定

5.1　データセット

実験には，3.1 章で述べた paraling-dial データセットから作成した感情的対話破綻検出データセットを用いた．4,470 組の音声・応答ペアから，感情と応答が一致する成立例と，不一致となる破綻例を 1:1 の比率で生成した．データセットは 8:1:1 に分割した．

5.2　比較手法の制御信号

提案手法では，感情ラベルを表す高レベルな one-hot ベクトルを制御信号として説明したが，その抽象度が性能に与える影響を検証するため，異なる抽象レベルを持つ制御信号を導入した比較モデルも

構築した．

- 低レベル抽象度の特徴量（音響特徴量）：MFCC [16], MFCCδ, MFCC$\delta\delta$, RMS エネルギー，基本周波数（F0），平均スペクトル重心，およびスペクトル重心の標準偏差
- 中レベル抽象度の特徴量（感情埋め込み）：上記の音響特徴量を入力とし，2 層ニューラルネットワークによる感情分類を学習させ，中間層の埋め込みベクトルを制御信号として利用した．この感情分類器の精度はテストデータにおいて 88.6%であった
- 高レベル抽象度の特徴量（one-hot ベクトル）：音響特徴量を入力として Random Forest [17] により感情分類を行い，その出力を one-hot ベクトルに変換した結果を使用した．この感情分類器は，テストデータにおいて 88.3%の分類精度を示した

全ての FiLM モデルの学習は共通のハイパーパラメータを使用した．最適化には Ada m（学習率 0.001），バッチサイズ 32，100 エポック（早打ち切りあり）で学習させた．

6 実験結果

本章では，提案手法の有効性を検証するとともに，制御信号として用いる情報の抽象度が性能に与える影響を分析し，得られた知見について考察する．

6.1 提案手法の有効性

まず，Qwen2-Audio を利用して提案する感情的対話破綻検出タスクを評価した．この結果，性能は 52.4%であり，既存の SLM では今回提案するような感情的対話破綻検出のタスクに対応する十分な音声特徴検出が行えていない可能性が示唆された．すなわち，既存の SLM が扱われているパラ言語情報が，発話のニュアンスやトーンを十分に扱うような情報を含んでいない，ということが示唆されている．次に，FiLM に基づく提案法により，感情をテキスト解釈を変調する制御信号として用いるという仮説の妥当性を検証した．高レベルの抽象度の特徴量として感情ラベルの正解ラベルを制御信号として与えた場合，FiLM モデルは 89.5%の精度で感情的対話破綻タスクを識別できた．提案アーキテクチャが感情的対話破綻を高精度に検出できることを確認できたと言える．

6.2 制御信号の抽象度による性能比較

次に，制御信号の情報抽象度が性能に与える影響を検証した．高レベル（感情ラベル one-hot），中レベル（感情埋め込み），低レベル（音響特徴量）の 3 種類の抽象度の特徴量を用いた場合の精度比較を表 1 に示す．

表 1 制御信号の抽象レベルと精度

モデル（制御信号）	精度
Qwen2-Audio(Few-shot)	52.4%
FiLM(高レベル：感情ラベル)	89.5%
FiLM(中レベル：感情埋め込み)	65.2%
FiLM(低レベル：音響特徴量)	47.7%

この結果から，感情的対話破綻検出タスクにおいては低レベルの音響特徴量が検出に寄与せず，より高レベルの感情ラベルに近い特徴量が精度向上に寄与することが明らかになった．つまり，パラ言語情報として含まれる発話のニュアンスやトーンを適切に捉えるには，より適切な抽象度の問題を合わせて解く必要がある，ということが示唆された．

7 結論

本研究では，音声対話において扱われるべき発話音声が持つパラ言語的なニュアンスを捉えることを指向して，感情的対話破綻のベンチマークデータセットの構築と，その破綻検出モデルの構築を行った．具体的には，同じ発話内容を異なる感情ラベルを与えて発話し，発話のニュアンスに応じて異なる応答を付与した paraling-dial データセットを構築した．このデータセットから感情的対話破綻検出タスクを定義し，感情がテキスト解釈を動的に変調する制御信号として扱われれる FiLM に基づく破綻検出器を構築した．FiLM を基盤としたモデルにより，89.5%の精度を達成し，仮定の妥当性と実用性を確認した．また，FiLM で用いるべき制御信号の抽象度について，感情的対話破綻検出のタスクにおいては，より感情ラベルに近い高次の特徴量表現が分類精度の向上に寄与することが明らかになった．これは，今後の SLM 学習において，タスクごとに関連度の高いタスクや表現を用いることの重要性を示唆している．

謝辞

本研究は，科研費 22H03654 と 22K17958 の支援を受けた．

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[2] Joshua J Guyer, Pablo Briñol, Thomas I Vaughan-Johnston, Leandre R Fabrigar, Lorena Moreno, and Richard E Petty. Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments. **Journal of nonverbal behavior**, Vol. 45, No. 4, pp. 479–504, 2021.

[3] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. Paralinguistics in speech and language— state-of-the-art and the challenge. **Computer Speech & Language**, Vol. 27, No. 1, pp. 4–39, 2013.

[4] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. **arXiv preprint arXiv:2407.10759**, 2024.

[5] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 29, pp. 3451–3460, 2021.

[6] Bilyana Martinovsky and David Traum. The error is the clue: Breakdown in human-machine interaction. 2006.

[7] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 32, 2018.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[9] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, 2004.

[10] Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, et al. Wavreward: Spoken dialogue models with generalist reward evaluators. **arXiv preprint arXiv:2505.09558**, 2025.

[11] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. **ACM computing surveys (CSUR)**, Vol. 47, No. 3, pp. 1–36, 2015.

[12] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. **Language resources and evaluation**, Vol. 42, pp. 335–359, 2008.

[13] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. **arXiv preprint arXiv:1810.02508**, 2018.

[14] James A Russell. A circumplex model of affect. **Journal of personality and social psychology**, Vol. 39, No. 6, p. 1161, 1980.

[15] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. **arXiv preprint arXiv:1908.10084**, 2019.

[16] Beth Logan, et al. Mel frequency cepstral coefficients for music modeling. In **Ismir**, Vol. 270, pp. 1–11. Plymouth, MA, 2000.

[17] Leo Breiman. Random forests. **Machine learning**, Vol. 45, No. 1, pp. 5–32, 2001.