パラ言語情報に着目した Speech-to-Text 対話ベンチマークデータセットの提案

中畔彪雅 [1]　河野誠也 [2,1]　Canasai Kruengkrai[2]　Angel Garcia Contreras[2]
千葉祐弥 [3]　杉山弘晃 [3]　吉野幸一郎 [4,2,1]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所
[3]NTT コミュニケーション科学基礎研究所　[4] 東京科学大学
nakaguro.hyuga.nl1@is.naist.jp
{seiya.kawano, canasai.kruengkrai, angel.garciacontreras}@riken.jp
{yuya.chiba, hiroaki.sugiyama}@ntt.com　koichiro.yoshino@riken.jp

概要

同じ発話文でもパラ言語情報が異なれば，与える意図やニュアンスが異なる．音声対話システムが持つこれらパラ言語情報処理能力を測るため，パラ言語情報の特に感情に着目した Speech-to-Text 対話ベンチマーク「paraling-data」を提案する．同じ発話文にパラ言語情報のみが異なる発話を収集し，それぞれのニュアンスに対応する応答を収集した．また，テキスト対話システムと Speech-to-Text 対話システムを構築し，感情ラベル予測を補助タスクとした対話応答生成を学習し評価した．

1　はじめに

人の発話が持つ細かな意図やニュアンスを汲み取ることは，今後の音声対話システムが持つべき重要な機能である．人の発話には言語的情報以外に韻律，発話速度，声量などのパラ言語情報が含まれ [1]，人は会話においてこれらの要素を巧みに操作しながら対話の細かいニュアンスを相手に伝達している．言い換えれば，発話文が同じでも，付与されているパラ言語情報が異なれば発話意図が変化する場合がある．

既存の音声対話システムは音声認識システムを大規模言語モデル（Large Language Model; LLM）などの対話モデルと接続することによって実現される．この際，ユーザー発話を音声から対話モデルの入力に適するテキストへと変換する段階で，これらパラ言語情報は失われてしまう．こうしたパラ言語情報を補完する目的で，音声をテキスト化せず離散的な整数列や埋め込み表現に変換するアプローチ（Speech-to-Text 対話システム）が存在する [2, 3, 4]．ただし，これらの枠組みで実際にどのようなパラ言語情報に含まれる意図がシステム応答に反映されたのかは明確ではない．

パラ言語情報に含まれる考慮するべき情報の代表として感情ラベルがある．既存の音声対話データセットにも感情ラベルが付与されたものは存在する [5, 6, 7]．ただし，Speech-to-Text 対話システムが持つパラ言語情報の処理能力を調査したい場合，言語的情報を固定した上で，パラ言語情報を変化させた場合にどうなるか確認できることが望ましい．

これらの問題を解決するために，本研究では全く新しい，パラ言語情報に着目した Speech-to-Text 対話ベンチマークである「paraling-data」を提案する．具体的には，全く同じ発話文（言語的情報）に対して異なる五つの感情を載せた音声を収録し，それらに対して感情情報も考慮したような応答を出力するタスクを設定する．

本研究の貢献は以下の通りである．

- 同じ発話文に異なるパラ言語情報を付与した発話とそれに対する応答を収集したデータセット「paraling-data」の構築
- テキスト対話システムと Speech-to-Text 対話システムで生成された応答を比較し，パラ言語情報が応答に与える影響を調査

2　パラ言語情報に着目したデータセット構築

本研究では，Speech-to-Text 対話システムのパラ言語情報処理能力，特に感情情報の処理能力を測ることに焦点を置く．既存の感情音声データセットでは，感情ラベルなどのパラ言語情報に応じて発話文

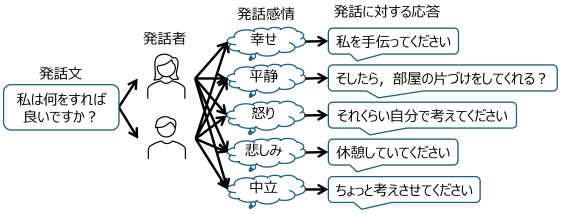

図 1　paraling-data の発話文および応答例

が変化している [5, 8, 6]．そのため，パラ言語情報のみが変化した発話に対する音声対話システムの処理能力を評価することは困難である．そこで，本研究では同じ発話文に対して二人の話者が五つの感情で発話し，それぞれの発話に対応する応答を収集した「paraling-data」を構築した（図 1）．

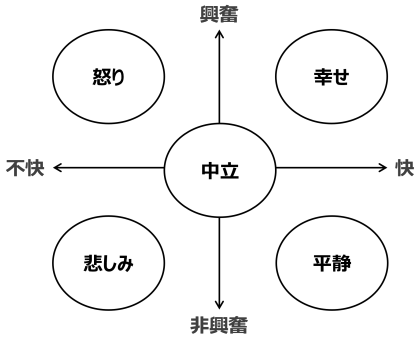

図 2　paraling-data の発話に含まれる感情設定

データセットの具体的な構築手順としては，話者が読み上げる短文を 149 種類収集した．それら発話文に対して話者（男性一名，女性一名）が五つの感情で発話することで，合計 1490 件の発話データを収集した．五つの感情はラッセルの感情円環モデル [9] の二軸である valence と arousal から感情（快，不快）と興奮（興奮，非興奮）を基に設定した．感情はそれぞれ幸せ（快・興奮），平静（快・非興奮），怒り（不快・興奮），悲しみ（不快・非興奮），中立である（図 2）．また，それぞれの発話に対する応答文をテキストで収集した．応答文は各発話文と各感情を考慮した発話に対して収集しており，性別は考慮していない．感情を考慮した応答が困難なものを除き，659 件の応答文を収集した．最終的に合計 1318 件のデータセットを構築した．

3　性能比較に用いるモデル

実際に構築した paraling-data がどの程度の難易度なのか検証する必要がある．そこで，テキスト対話モデルと Speech-to-Text 対話モデルを構築し，それぞれの paraling-data に対する応答の精度や多様性を比較する．

3.1　テキスト対話システム

日本語対話で事前学習された “rinna-japanese-gpt-neox-3.6b-instruction-ppo”[10] を用いる．入力発話の書き起こしテキストである発話文を入力し，応答を生成する．

プロンプトは以下の通りである．

<s>ユーザー：{発話文}<NL>[CLS] システム：{応答}</s>

“< s >” と “< $/s$ >” は文の始まりと終わりを表す特殊トークンである．“< NL >” は改行であり，ここではユーザーとシステムのターン切り替え位置を表している．“[CLS]” は感情予測の際に使用するトークンである．

3.2　Speech-to-Text 対話モデル

音声エンコーダ・アダプター・k-means クラスタ・対話モデルで構成する．音声エンコーダとアダプターには，日本語対応の音声認識システムである Nue-ASR[11] の HuBERT[12]，二層の畳み込み層と一層の全結合層で構成されたものを用いる．音声エンコーダは音声を 768 次元の音響特徴量系列に変換する．音響特徴量系列をアダプターで時間軸方向へ 1/4 に圧縮し，次元数を 2816 次元に拡張する．k-means クラスタはアダプターから得られた特徴量を 1000 クラスに分類し，音声を離散的な整数列に変換する．使用する対話モデルの入出力は本来テキストだが，この処理により音声をテキストと同様にトークン化して対話モデルへ入力する．対話モデルは，節 3.1 と同様の物を用いる．

プロンプトは以下の通りである．

<s>ユーザー：{音声トークン列}[SEP]{発話文}<NL>[CLS] システム：{応答}</s>

“[SEP]” は音声トークン列と発話文の区切りを表す．

3.3　補助タスク：感情予測

Speech-to-Text 対話システムがテキスト対話システムよりもパラ言語情報を考慮した応答を出力することを期待する．ただし，今回のベンチマークは発話が持つ感情状態によって応答が変換するため，感情ラベルを明示的に扱う方が望ましい．そこで，対話タスクに感情予測を補助タスクとして追加する．

対話応答生成時に特殊トークン [CLS] を出力し，

このトークンの隠れ状態から一層の全結合層で感情を予測する．感情予測の損失 L_e は対話応答生成の損失 L_r に加算され，全体の損失 $L_e + L_r$ を最小化させるよう学習する．

4　感情予測を補助タスクとした対話応答生成学習

4.1　音声適応事前学習

事前学習された LLM はテキストで学習されており，Speech-to-Text 対話システムの構築には音声に適応する学習が必要である．また，後の感情予測に使用する特殊トークン “[CLS]” を含めて学習を行う．そこで，雑談音声対話データセットの NUCC[13] を節 3.2 で示したプロンプトの形式で学習する．

データセットは学習: 検証: テスト=9:0.5:0.5 に分割する．バッチサイズ=32，エポック数=1，学習率=5e-5，最適化手法には AdamW を用いて，LLM の全パラメータを更新する．損失関数には交差エントロピー誤差を用いる．最終的なモデルは，総学習ステップ数の各 20 %終了時の検証データに対する損失が最も低いものを選択する．また，選択したモデルのテストデータに対する損失は 2.041 であった．

パラ言語情報処理能力の評価ではテキスト対話システムと比較を行うため，同条件でテキスト対話システムの対話モデルも学習する．プロンプトは節 3.1 で示した形式である．テストデータに対する損失は 1.615 であった．

4.2　感情を考慮する対話応答ファインチューニング

感情ラベルを明示的に扱うために感情予測を補助タスクとして追加する．感情の予測は出力される [CLS] の隠れ状態から一層の全結合層を用いて行い，対話応答生成と共に学習する．対話応答生成に加えて感情予測が必要であるため，学習するデータセットには感情ラベルが付与されている音声対話データの STUDIES[5] と構築した paraling-data の一部を混合したものを使用する．paraling-data で使用するデータは，感情を考慮した応答が困難で五感情の全てに対して応答が用意されていない 318 件を用いる．

STUDIES データは学習: 検証: テスト=8:1:1 に分割し，学習データと検証データに paraling-data の 318 件を 1:1 に分割し加える．バッチサイズ=32，エポック数=3，学習率=5e-5，最適化関数には AdamW を用いて，対話モデルである LLM の Attention 層に適応した LoRA（$r = 8, \alpha = 16$）[14] と感情予測のための全結合層を学習する．損失関数は生成された応答に対する交差エントロピー誤差 L_r と感情予測に対する交差エントロピー誤差 L_e を用いて，$L_r + L_e$ とする．最終的なモデルは，各エポック終了時の検証データに対する損失が最も低いものを選択する．また，事前学習と同様にテストデータで評価した結果，テキスト対話システムの損失は 3.191 であり，Speech-to-Text 対話システムの損失は 3.331 となった．

5　パラ言語情報処理能力評価

テキスト対話モデルと Speech-to-Text 対話モデルの paraling-data に対する応答と感情予測からパラ言語情報に対する処理能力を評価する．評価に用いるデータは学習に使用していない五感情の全てに応答が用意されているデータで，各感情ごとに 200 件あり合計 1000 件のデータとなる．発話に含まれるパラ言語情報が応答に与える影響を評価するために，生成確率が最も高いトークンを選択していく Greedy 法により生成した．

生成された応答に対する自動評価項目は BLEU-4 と distinct-1,2 である．BLEU-4 で正解データとの一致割合，distinct-1 で生成単語の多様性，distinct-2 で生成文法の多様性を評価する．加えて，各応答を日本語で事前学習された Sentence-BERT[15] で固定長ベクトルに変換し，t-SNE[16] を用いて二次元にマッピングした分布および各モデルの感情予測精度からパラ言語情報が応答に与える影響を見る．

表 1　BLEU および distinct，感情予測精度

	BLEU-4	dist-1	dist-2	感情予測精度（%）
テキスト	0.128	0.009	0.019	20
Speech-to-Text	0.118	0.019	0.046	21.5
正解応答		0.072	0.216	

各モデルが生成した応答に対する BLEU-4 と正解応答を含めた distinct-1,2 を表 1 に示す．テキスト対話システムはより正解データに近い応答を生成し，Speech-to-Text 対話システムはより多様な応答を生成した．しかし，感情予測精度が非常に低く，応答の生成に感情を考慮することはできていない可能性が高い．

各モデルが生成した応答文と正解の応答文を Sentence-BERT と t-SNE により二次元にマッピング

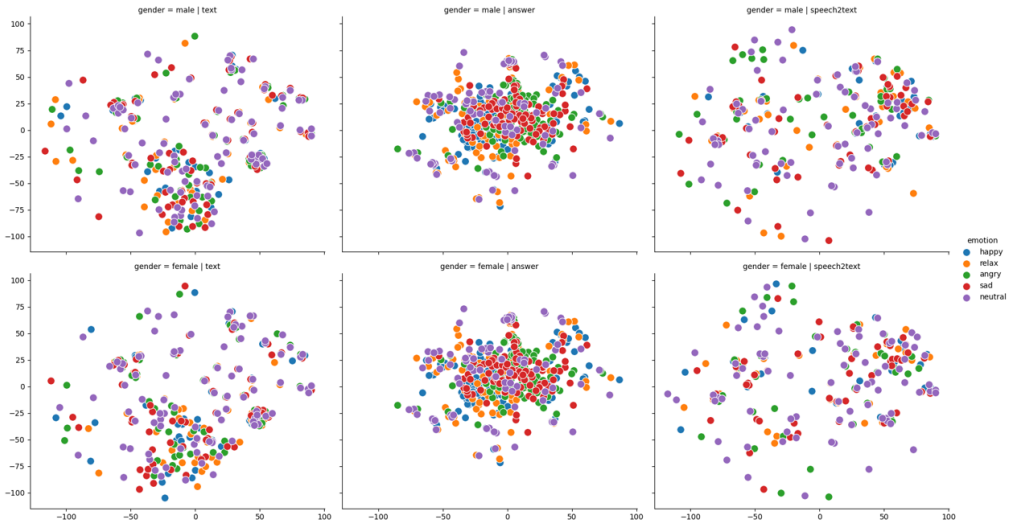

図 3 各モデルが生成した応答文と正解データの散布図
（左：テキスト対話システム，中央：正解，右：Speech-to-Text 対話システム）

した散布図を図 3 に示す．上段は男性発話者，下段は女性発話者，各色は各感情の発話を表し，各点はそれぞれの要素を含む発話に対する応答文を投影したものである．各モデルが生成した応答は正解データよりも意味的に多様であることが分かるが，話者や感情ごとに偏りは見られなかった．

テキスト対話システムは平静のみ，Speech-to-Text 対話システムは幸せと平静のみを予測感情として生成していた．節 4.2 の感情予測を補助タスクとした対話応答生成学習に用いた学習データセットの発話感情は幸せ：912, 平静：1550, 怒り：496, 悲しみ：708, 中立：45 件であり，幸せと平静に偏っていた．これらから学習データセットに含まれる感情ラベルの偏りにより，感情予測精度が低くなったと考えられる．

応答が感情を捉えておらず，テキスト対話システムと Speech-to-Text 対話システムでパラ言語情報に対する応答の変化に差が出ず，パラ言語情報に対する処理能力を評価できなかった．能力評価にはパラ言語情報の要素に対する予測精度向上が必要である．また，本研究で使用した評価指標では当初の目的であるパラ言語情報に対する処理能力を直接的に評価するには不十分であり，構築したベンチマークに合わせてパラ言語情報に対する処理能力を評価する新たな指標が必要である．

6 まとめ

既存の音声対話データセットは感情ラベルに応じて発話文が変化しており，パラ言語情報に着目した処理能力を評価することは困難である．そこで，本研究では同じ発話文に異なるパラ言語情報を付与した音声データと，それぞれに対応した応答テキストを収集した「paraling-data」を構築した．また，テキスト対話システムと Speech-to-Text 対話システムを構築し，paraling-data に対する応答の変化からパラ言語情報に対する処理能力を評価した．結果，感情予測精度が低く感情を考慮した応答生成を確認できず，パラ言語情報の処理能力を評価できなかった．パラ言語情報に対する処理能力を評価するためには，応答生成能力だけでなくパラ言語情報の要素に対する予測精度の向上が必要である．今後は更なるデータセットの拡張およびパラ言語情報の処理能力を評価する新たな指標の開発に取り組む．

謝辞

本研究では，国立情報学研究所 音声資源コンソーシアムから提供を受けた「日本語共感的音声対話コーパス（STUDIES）」を利用した．本研究は，科研費 23K24910 と 22K17958 の支援を受けた．

参考文献

[1] 藤崎博也. 音声の音調的特徴のモデル化とその応用. 文部省科学研究費特定領域研究 「韻律に着目した音声言語情報処理の高度化」 研究成果報告書, 2005.

[2] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. **arXiv preprint arXiv:2305.11000**, 2023.

[3] Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. Llasm: Large language and speech model. **arXiv preprint arXiv:2308.15930**, 2023.

[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] 高道 慎之介 橘 健太郎 猿渡 洋齋藤 佑樹. Studies：表現豊かな音声合成に向けた日本語共感的対話音声コーパス. 日本音響学会 2022 年春季研究発表会 講演論文集, 2-3P-15, pp. 1133–1136, 2022 年 3 月.

[6] 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.

[7] Patrick S Kamath and W Ray Kim. The model for end-stage liver disease (meld). **Hepatology**, Vol. 45, No. 3, pp. 797–805, 2007.

[8] Kazuya Saeki, Masaharu Kato, and Tetsuo Kosaka. Language model adaptation for emotional speech recognition using tweet data. In **2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)**, pp. 371–375, 2020.

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

[10] Tianyu Zhao and Kei Sawada. rinna/japanese-gpt-neox-3.6b-instruction-ppo.

[11] Yukiya Hono, Koh Mitsuda, Tianyu Zhao, Kentaro Mitsui, Toshiaki Wakatsuki, and Kei Sawada. An integration of pre-trained speech and language models for end-to-end speech recognition. **arXiv preprint arXiv:2312.03668**, 2023.

[12] 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.

[13] 藤村逸子, 大曽美恵子, 大島ディヴィッド義和. 会話コーパスの構築によるコミュニケーション研究. 藤村逸子、滝沢直宏編『言語研究の技法：データの収集と分析』p. 43-72、ひつじ書房, 2011.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.

[15] 園部勲. sonoisa/sentence-bert-base-ja-mean-token-v2.

[16] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. **Journal of machine learning research**, Vol. 9, No. 11, 2008.