音声言語モデルに基づくパーソナライズド音声感情認識のための効果的な In-Context Lening の検討

庵 愛　山田 涼楓　山根 大河　牧島 直輝　田中 智大　鈴木 聡志　折橋 翔太　増村 亮
NTT 株式会社　人間情報研究所
mana.ihori@ntt.com

概要

音声感情認識（Speech Emotion Recognition: SER）タスクでは，音声言語モデルにおいて対象話者の発話とその感情ラベルの事例を用いた in-context learning（ICL）を実施することで，対象話者へのパーソナライズが可能となることが報告されている．従来，大規模言語モデルや音声認識タスクにおける ICL では事例の提示順やその質が性能に影響を与えることが明らかにされている．一方，SER タスクのパーソナライゼーションに ICL を活用する場合は，事例のどのような要素がどのような影響を与えるかが明らかではない．そこで本稿では，音声言語モデルに基づくパーソナライズド SER のための効果的な ICL を行うための要素を明らかにする．

1　はじめに

音声感情認識（Speech Emotion Recognition: SER）は，人間とコンピュータ，あるいは人間同士のコミュニケーションを理解するために重要な役割を担っている．例えば，コンタクトセンタにおける顧客の満足度推定 [1] や，同情や共感のできる音声対話システムへの応用 [2] が期待されている．しかし，感情の表出傾向は文化や環境，個人の特性によって大きく異なるため，SER は一般化が難しい [3]．

未知話者への SER 性能を向上させるため，対象話者の登録発話を推論時に条件づけることで SER モデルをその話者にパーソナライズする方法がある [4, 5]．登録発話とは，事前に録音された対象話者の感情状態を表す発話のことであり，モデルはこの発話を手掛かりに SER をパーソナライズする．ここで，登録発話の数や感情の種類を柔軟に調整するために，大規模言語モデル（Large Language Model: LLM）を音声入力が可能となるように拡張した音声言語モデル（Speech Language Model: SLM）における in-context learning（ICL）[6] を活用したパーソナライゼーション手法が提案されている [7]．この手法では，ICL に用いる登録発話とそのラベルの数が増加するほど性能が向上し，動的なパーソナライゼーションが可能となった．しかし，ここで示された結果は全感情の平均精度のみであり，登録発話の選択方法のバリエーションも限定されたものであった．

ICL の活用に関して，LLM では，事例の選択や提示順によって性能が大きく変動することが報告されている [8, 9]．また，音声言語モデルにおける ICL を音声認識タスクに活用した場合，事例に用いる音声サンプルの選択が性能に影響することが報告されている [10]．そのため，ICL を SER タスクのパーソナライゼーションに活用する場合にも，登録発話の特徴によって性能やその効果に影響を与える可能性がある．これらの特徴を明らかにすることで，SLM に基づくパーソナライズド SER において効果的な ICL を実施できるようになると考える．

そこで本稿では，SLM に基づくパーソナライズド SER において，より効果的な ICL を実施するために必要な要素を明らかにすることを目的とする．ICL に基づいて SER をパーソナライズする場合の特有の課題として，登録発話における感情の種類の選択方法が挙げられる．例えば，複数の登録発話において感情の重複がある場合とない場合で性能が変化するのかといったことである．また，ICL で共通した課題である，例に用いる音声サンプルやその提示順が与える影響や，各感情や各話者に対するパーソナライズ効果についても詳細な分析を行う．本稿で明らかになった，SLM に基づくパーソナライズド SER に対する効果的な ICL の特徴とその効果を以下に示す．

- 登録発話の感情の種類の選択では，目的発話と同じ種類の感情が含まれており，感情の種類の被りがないほど性能が向上する．

- パーソナライゼーションの効果はほとんどの話者に対して有益であり，劣化するケースはわずかである．

2 SLM の ICL に基づくパーソナライズド SER

本手法では，SLM に，指示テキスト $\boldsymbol{W}$，目的発話の音響特徴量 $\boldsymbol{X}$，同一話者の k 個の登録発話の音響特徴量とその感情ラベルを表すテキストのペア $\boldsymbol{S} = \{(\boldsymbol{X}^1, \boldsymbol{Z}^1), \cdots, (\boldsymbol{X}^k, \boldsymbol{Z}^k)\}$ が与えられたとき，目的発話の感情ラベルを表すテキスト $\boldsymbol{Z}$ の事後確率 $P(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{X}; \boldsymbol{\Theta})$ を予測する．ここで，$\boldsymbol{\Theta}$ は学習可能なパラメータセットである．SLM は，音声エンコーダ，LLM と音響特徴量の埋め込み空間を揃える変換モジュールの Q-Former[11]，テキストエンコーダ，LLM エンコーダ，LLM デコーダで構成される．モデルの概要を図 1 に示す．

学習方法　ICL に基づいて SER をパーソナライズするため，SLM は以下のように学習される．まず，パーソナライゼーションに用いる登録発話とそのラベルのペア数 k は $\{0, 1, \cdots, K\}$ から一様ランダムに決定される．ここで，K はデータセットに含まれる感情のカテゴリ数を表す．次に，$\boldsymbol{\Theta}$ は J 人の話者を含むデータセット $\mathcal{D} = \{\mathcal{D}_1, \cdots, \mathcal{D}_j, \cdots, \mathcal{D}_J\}$ を用いて式 (1) によって最適化される．ここで，j 番目のデータセットは $\mathcal{D}_j = \{(\boldsymbol{X}_j^1, \boldsymbol{Z}_j^1), \cdots, (\boldsymbol{X}_j^{|\mathcal{D}_j|}, \boldsymbol{Z}_j^{|\mathcal{D}_j|})\}$ で構成される．

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} - \sum_{j=1}^{J} \sum_{(\boldsymbol{X}_j, \boldsymbol{Z}_j) \in \mathcal{D}_j} \begin{cases} \log P(\boldsymbol{Z}_j|\boldsymbol{W}, \boldsymbol{X}_j; \boldsymbol{\Theta}) & \text{if } k = 0, \\ \log P(\boldsymbol{Z}_j|\boldsymbol{W}, \boldsymbol{S}_j, \boldsymbol{X}_j; \boldsymbol{\Theta}) & \text{otherwise.} \end{cases} \tag{1}$$

$$\boldsymbol{S}_j = \texttt{SelectProcedure}(\boldsymbol{X}_j, \boldsymbol{Z}_j, k, \mathcal{D}_j) \tag{2}$$

SelectProcedure() は $\mathcal{D}_j$ から $(\boldsymbol{X}_j, \boldsymbol{Z}_j)$ を除く k 個の $\boldsymbol{X}$ と $\boldsymbol{Y}$ のペアを被りがないように一様ランダムに選択する．また，$k = 0$ のときはパーソナライズしない SER が学習される．

ICL に基づくパーソナライゼーション　SLM は，k 個の対象話者の登録発話とそのラベルのペアを条件づけることにより，ICL によって SER をパーソナライズする．k-shot の ICL では，$\boldsymbol{Z}$ は式 (3) によって推論される．

$$\hat{\boldsymbol{Z}} = \arg\max_{\boldsymbol{Z}} \begin{cases} P(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{X}_{\text{tgt}}; \hat{\boldsymbol{\Theta}}) & \text{if } k = 0, \\ P(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{S}_{\text{tgt}}, \boldsymbol{X}_{\text{tgt}}; \hat{\boldsymbol{\Theta}}) & \text{otherwise.} \end{cases} \tag{3}$$

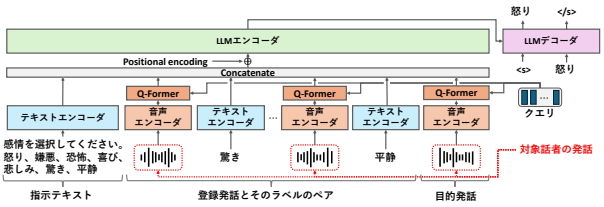

図 1　SLM の ICL に基づくパーソナライズド SER の概要

ここで，$\boldsymbol{S}_{\text{tgt}} = \{(\boldsymbol{X}_{\text{tgt}}^1, \boldsymbol{Y}_{\text{tgt}}^1), \cdots, (\boldsymbol{X}_{\text{tgt}}^k, \boldsymbol{Y}_{\text{tgt}}^k)\}$ を表し，$(\boldsymbol{X}_{\text{tgt}}^k, \boldsymbol{Y}_{\text{tgt}}^k)$ は対象話者の k 番目の登録発話とそのラベルのペアを示す．また，$k = 0$ のとき，SLM はパーソナライズせずに SER を実行する．

3 評価実験

データセット　本稿では，多様な未知話者で調査を実施するために，[7] と同様のデータセットを採用する．本データセットは，10 から 70 代の男性 368 名，女性 432 名の日本語を母国語とする話者の 7 つの感情（怒り，嫌悪，恐怖，喜び，悲しみ，驚き，平静）の演技音声を，各感情 50 発話ずつ録音したものである．本データセットを話者数，発話数が学習セットで 643 人と 225,050 件，開発セットで 79 人と 27,650 件，評価セットで 78 人と 27,300 件となるように分割した．また，指示テキストは付録に示す．

モデル　SLM では，我々の作成した 0.6B のパラメータを持つ transformer encoder-decoder 型の LLM を用いる．この LLM は unifying language learning paradigms[12] に基づいて大量の日本語テキストデータで事前学習した後，様々な自然言語処理タスクで instruction-tuning したものである．また，音声エンコーダは我々の作成した 42M のパラメータをもつ transformer ベースのエンコーダを用いる．この音声エンコーダは，様々な日本語の音声理解タスクや音声認識によって事前学習されたものである．

実験設定　音響特徴量は 80 次元のログメルフィルタバンク特徴量を用い，フレームシフトは 10ms とした．音声エンコーダ，Q-Former，LLM エンコーダ，デコーダをそれぞれ 512，512，1024，1024 次元のユニットと 2048，2048，4096，4096 次元の position-wise feed forward network を持つ 6 層，2 層，32 層，6 層の transformer block で構成した．Q-Former のクエリの次元数は予備実験で最高性能を達成した 150 に設定した．学習には RAdam を用い，ミニバッチサイズは 64 に設定した．推論にはビーム幅 4 のビームサーチアルゴリズムを採用し，ICL では 0 から 7 個の登録発話とラベルのペアをそれぞれ条件

表 1 各選択方法を用いた場合の全ての感情に対する UA（%）

shot	A	B	C	D	E 全て怒り	全て嫌悪	全て恐怖	全て喜び	全て悲しみ	全て驚き	全て平静
0	66.4	66.4	66.4	66.4	66.4	66.4	66.4	66.4	66.4	66.4	66.4
1	69.8	69.8	69.2	69.2	69.5	69.6	70.4	69.9	70.2	70.3	69.0
2	71.5	71.7	70.8	70.6	70.5	70.3	71.2	71.1	70.9	71.1	69.6
3	72.6	73.2	71.4	71.8	70.9	70.5	71.7	71.6	71.1	71.6	70.0
4	73.4	73.8	72.1	72.6	71.2	70.8	71.6	72.1	71.0	72.0	69.9
5	74.0	74.7	72.7	73.4	–	–	–	–	–	–	–
6	74.4	75.2	72.7	74.0	–	–	–	–	–	–	–
7	74.9	75.7	73.4	–	–	–	–	–	–	–	–

づけた．評価指標には，話者ごとの非加重平均精度（Unweight Accuracy: UA）を採用する．ここで，SLM が生成したテキストが参照データと完全一致した場合に正解とした．

4 分析

4.1 登録発話における感情の選択方法

登録発話の選択のうち，感情の種類が性能に与える影響を調査するため，以下の 5 つの感情の選択方法について性能を比較し，パーソナライズド SER のための ICL で効果的な感情の組み合わせを明らかにする．**A：一様ランダムに選択．B：感情の種類がコンテキスト内で重複しないように一様ランダムに選択．C：目的発話と同じ感情は含まないように，一様ランダムに選択．D：目的発話と同じ感情は含まないように，感情の種類がコンテキスト内で重複しないように一様ランダムに選択．E：同一感情の音声サンプルを一様ランダムに選択**．C，D の選択方法では，恣意的に目的発話と同じ登録発話を用いないことで，本手法では目的発話と同様の感情を表す登録発話を用いることが有効なのか，ICL を活用することが有効なのかを調査することを目的とする．

これらの結果を表 1 に示し，表中の値は全ての感情に対する平均の UA を示す．表 1 より，A から D の選択方法では，全ての方法で登録発話の数が増加するほど性能が向上しており，性能は B，A，D，C の順で向上した．一方，E の選択方法では，感情の種類によっては，登録発話の数が増加しても性能は飽和することが確認された．これらの結果より，目的発話と同じ感情を登録発話として用いることは有効ではあるが，用いない場合でも ICL によるパーソナライズ効果を獲得できることが示唆された．また，感情の種類をコンテキスト内で重複しない様に選択することで性能が向上することも示された．これは，様々な感情を使用することで，話者の感情の表出傾向を獲得できるためだと考えられる．以上より，B の選択方法を用いることでパーソナライズ効果を最大化できることが示された．

表 2 音声サンプルを変更したときの感情ごとの平均の UA（%）とその SD

	怒り	嫌悪	恐怖	喜び	悲しみ	驚き	平静
UA	79.2	77.6	63.2	71.1	73.1	71.6	92.3
SD	0.4	0.2	0.3	0.4	0.5	0.2	0.2

表 3 提示順を変更したときの感情ごとの平均の UA（%）とその SD

	怒り	嫌悪	恐怖	喜び	悲しみ	驚き	平静
UA	80.2	78.1	64.3	72.0	72.9	70.2	91.8
SD	0.5	0.5	0.8	0.8	1.0	1.2	0.2

4.2 登録発話の音声サンプルの選択方法やその提示順

ここでは，4.1 節の B の方法で選択した登録発話を用いた 7-shot 推論のうち，各感情における音声サンプルの選択やその提示順が性能に与える影響を調査する．まず，各感情の音声サンプルの変更が与える影響を調査するため，音声サンプルを一様ランダムに変更した推論を 10 回行い，感情ごとに平均の UA とその標準偏差（Standard Deviation: SD）を算出する．変更される音声サンプルは同一話者の同一の感情状態を示すものであるが，発話内容や発話の長さが異なる．ここで，音声サンプルの変更のみの影響を調査するため，感情の提示順は変更しない．次に，提示順の変更が与える影響を調査するために，登録発話の提示順を一様ランダムに変更した推論を 10 回行い，感情ごとに平均の UA とその SD を算出する．ここで，提示順のみの影響を調査するため，各感情で選択される音声サンプルは変更しない．

これらの結果を表 2，3 に示す．表 2 より，音声サンプルを変更した場合，SD は 0.5pt 以下に収まっており，ほとんど性能変動がないことがわかる．一方，表 3 より，提示順を変更した場合，恐怖，喜び，悲しみ，驚きの感情で SD が 1pt 程度となっており，

表 4　ICL の shot 数を変更したときの感情ごとの UA（%）

shot	怒り	嫌悪	恐怖	喜び	悲しみ	驚き	平静
0	75.9	66.3	42.2	65.1	66.3	60.0	88.9
1	79.0	70.3	48.0	65.9	71.3	64.7	89.5
2	79.5	72.7	52.5	66.7	73.2	66.8	90.4
3	80.5	75.5	56.3	68.0	73.4	68.6	90.3
4	79.8	74.8	59.3	68.3	73.8	69.3	90.8
5	79.9	76.0	60.7	70.8	74.7	69.1	91.4
6	80.9	77.2	62.5	71.8	74.0	69.0	91.1
7	80.3	77.9	64.2	72.0	73.6	69.7	91.8

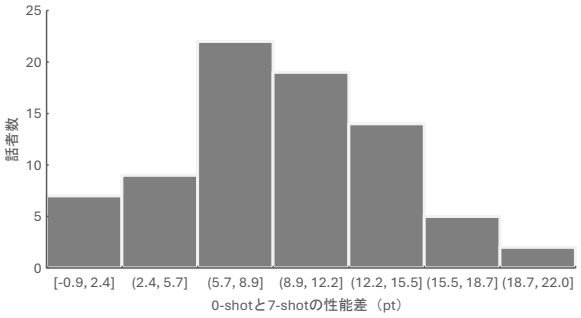

図 2　0-shot と 7-shot の性能差のヒストグラム

性能変動があることがわかる．これらの結果より，登録発話の提示順は性能に影響があるものの，その言語的な内容や発話の長さは性能に影響がない可能性が示された．提示順に関して，各感情で性能が高い上位 2 つの提示順を確認したところ，平静と怒りを除く全ての感情で，1，2 番目に目的発話と同様の感情を表す登録発話が提示されていた．そのため，認識したい感情と同様の感情を表す登録発話がある場合は，最初に条件づけることで効果的なパーソナライゼーションが実施できる可能性が考えられる．

4.3　各感情でのパーソナライズ効果

ここでは，4.1 節で明らかになった最適な感情の選択方法におけるパーソナライズ効果が，各感情でも有益であるかを調査する．具体的には，4.1 節の B の方法で ICL の登録発話を選択し，登録発話が増加した場合の感情ごとの性能の変化を調査する．その結果を表 4 に示す．表 4 より，全ての感情において，登録発話の数が増加するほど，性能が向上している．ここで，怒りと平静の感情では他の感情と比較して，2-shot 以降の性能向上が小さいことがわかる．これは，0-shot 時点の性能が他の感情と比較して大きいため，得られるパーソナライズ効果が小さくなったためだと考えられる．これらの結果より，全ての感情で ICL に基づくパーソナライズ効果を得ることができるが，その効果の大小は 0-shot 時点の

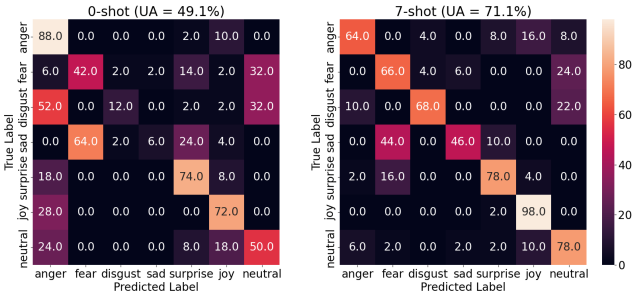

図 3　性能差が最大の話者の 0-shot と 7-shot の混同行列

認識性能によって異なることが示唆された．

4.4　各話者でのパーソナライズ効果

ここでは，4.1 節で明らかになった最適な感情の選択方法におけるパーソナライズ効果が，各話者でも有益であるかを調査する．そこで，テストセットの 78 名の各話者に対するパーソナライズしない 0-shot 推論と，4.1 節の B の方法で選択した登録発話を用いた 7-shot 推論の全ての感情に対する UA の平均値を算出し，パーソナライズした 7-shot とパーソナライズしない 0-shot の UA の値の差を算出する．その性能差のヒストグラムを図 2 に示す．図 2 より，パーソナライゼーションによって一話者で 1pt 程度性能が低下したが，その他の話者では性能が向上した．特に，2.4pt 以上性能が向上した話者は全体の 94%となり，ほとんどの話者で ICL によるパーソナライズ効果があることが明らかになった．

ここで，性能差が最大の話者の 0 と 7-shot 推論の混同行列を図 3 に示す．図中の UA の値は全ての感情に対する平均値を示し，混同行列の各行は 100%で正規化されている．図 3 より，0-shot では，ほとんどの発話を怒りに分類してしまっているが，パーソナライゼーションを実施することで怒りの誤認識を大幅に減らし，多くの感情を正しく認識できるようになっている．性能差がマイナス，最小の話者の混同行列の比較については付録に示す．

5　おわりに

本稿では，SLM に基づくパーソナライズド SER に効果的な ICL についての検討を行った．検討の結果，ICL における登録発話を，目的発話と同じ種類の感情を含み，感情の被りがないように選択することでパーソナライズ効果を高めることができ，その効果はほとんどの話者に対して有益であることを明らかにした．

参考文献

[1] Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, Yushi Aono, and Tomoki Toda. Customer Satisfaction Estimation in Contact Center Calls Based on a Hierarchical Multi-Task Model. **IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLPRO)**, pp. 715–728, 2020.

[2] Jaime C. Acosta. Using Emotion to Gain Rapport in a Spoken Dialog System. In **Proc. North American Chapter of the Association for Computational Linguistics (NAACL)**, p. 49–54, 2009.

[3] Ryne A Sherman, John F Rauthmann, Nicolas A Brown, David G Serfass, and Ashley Bell Jones. The Independent Effects of Personality and Situations on Real-time Expressions of Behavior and Emotion. **Journal of personality and social psychology**, p. 872, 2015.

[4] Andreas Triantafyllopoulos, Shuo Liu, and Björn W. Schuller. Deep speaker conditioning for speech emotion recognition. In **Proc. IEEE International Conference on Multimedia and Expo (ICME)**, pp. 1–6, 2021.

[5] Andreas Triantafyllopoulos and Björn Schuller. Enrolment-based personalisation for improving individual-level fairness in speech emotion recognition. In **Proc. Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 3729–3733, 2024.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In **Proc. Advances in Neural Information Processing Systems (NeurIPS)**, pp. 1877–1901, 2020.

[7] Mana Ihori, Taiga Yamane, Naotaka Kawata, Naoki Makishima, Tomohiro Tanaka, Satoshi Suzuki, Shota Orihashi, and Ryo Masumura. Few-shot Personalization via In-Context Learning for Speech Emotion Recognition based on Speech-Language Model. In **Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)**, 2025.

[8] Shengnan An, Bo Zhou, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Weizhu Chen, and Jian-Guang Lou. Skill-Based Few-Shot Selection for In-Context Learning. In **Proc. the Empirical Methods in Natural Language Processing (EMNLP)**, pp. 13472–13492, 2023.

[9] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proc. the Association for Computational Linguistics (ACL)**, pp. 1423–1436, 2023.

[10] Nathan Roll, Calbert Graham, Yuka Tatsumi, Kim Tien Nguyen, Meghan Sumner, and Dan Jurafsky. In-Context Learning Boosts Speech Recognition via Human-like Adaptation to Speakers and Language Varieties. **arXiv preprint arXiv:2505.14887**, 2025.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In **Proc. International Conference on Machine Learning (ICML)**, pp. 19730–19742, 2023.

[12] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. Ul2: Unifying language learning paradigms. In **Proc. International Conference on Learning Representations (ICLR)**, 2023.

A　指示テキスト

本稿では，全ての場合において以下の指示テキストを共通して用いた．

> 入力音声の感情を次の候補の中から適切なものを選んでください。
> - 平静の感情の音声です。
> - 驚きの感情の音声です。
> - 悲しみの感情の音声です。
> - 喜びの感情の音声です。
> - 恐怖の感情の音声です。
> - 嫌悪の感情の音声です。
> - 怒りの感情の音声です。
>
> なお、各感情音声は発話内容に関係なく、以下の特徴を持ちます。
> - 平静 = 淡々とした声
> - 驚き = びっくりした様子が伝わるイントネーションが付いた声
> - 悲しみ = 小さめの気落ちした声
> - 喜び = 嬉しさがこもったイントネーションが付いた声
> - 恐怖 = 押し殺したような声
> - 嫌悪 = 軽蔑が伝わる声
> - 怒り = 語気が荒い声

そのため，例えば目的発話の感情が怒りだった場合，出力の感情テキストは以下のように生成される．

> 怒りの感情の音声です。

B　パーソナライズ効果が小さい話者の特徴

ここでは，図 2 の性能差がマイナスであった話者 A と最小の話者 B における混同行列を分析することで，音声言語モデルに基づくパーソナライズド音声感情認識のための in-context learning の効果が小さい話者の特徴を明らかにする．図 4 に話者 A のパーソナライズを行わない 0-shot 推論と 4.1 節の B の方法で登録発話を選択した 7-shot 推論の各混同行列を，図 5 に話者 B の各混同行列を示す．各図の UA の値は全ての感情に対する平均値を示し，混同行列の

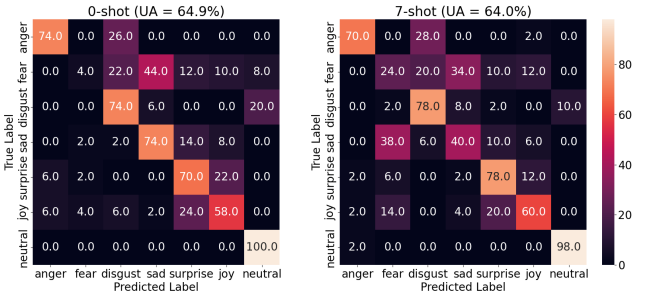

図 4　話者 A の 0-shot と 7-shot の混同行列

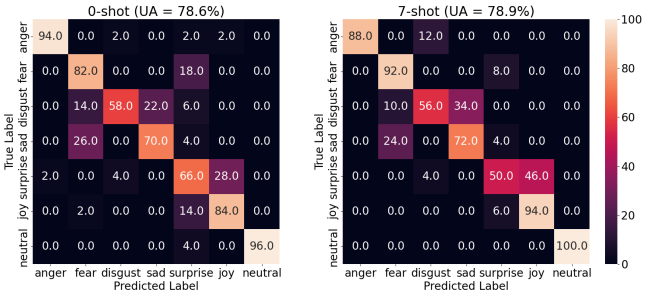

図 5　話者 B の 0-shot と 7-shot の混同行列

各行は 100%で正規化されている．まず，図 4 より，全ての感情に対する UA の値は低下しているものの，恐怖の感情では値が 20pt 向上している．一方，7-shot の結果では，悲しみと恐怖が混同されることが多く，悲しみの感情では値が 34pt 低下している．また，図 5 に関しても，全ての感情に対する UA の値の変化は小さいものの，恐怖と喜びの感情では値が 10pt 向上している．一方，7-shot の結果では，驚きが喜びに混同されることが多く，驚きの感情では値が 16pt 低下している．これらの結果より，表出が似ている感情がある場合，それらの感情を ICL に用いるとどちらか，または両方の感情の性能を低下させる可能性が示された．