大規模音声認識モデルに基づく韻律・言語情報を考慮した音声感情認識

福田りょう[1]　叶高朋[1]　安藤厚志[2]　小川厚徳[1]
[1]NTT コミュニケーション科学基礎研究所　[2]NTT 人間情報研究所
{ryo.fukuda,takatomo.kanou,atsushi.ando,atsunori.ogawa}@ntt.com

概要

本研究では、大規模音声認識モデルである Whisper に基づく音声感情認識手法を提案する。近年盛んに研究が行われている、事前学習済み音声エンコーダに基づく音声感情認識手法では、言語的な情報を十分に考慮することができなかった。そこで多量のデータ・複数の音声言語処理タスクで事前学習された Whisper のデコーダを用いることで、言語情報も考慮した感情認識手法の実現を目指した。また、音声認識や性別認識といったサブタスクのトークンを含めた単一系列をデコーダで生成するマルチタスク学習手法を提案した。

1　はじめに

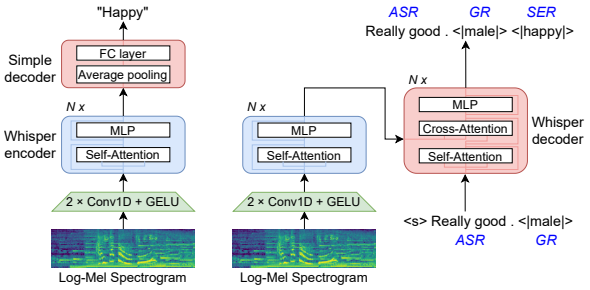
音声感情認識（Speech Emotion Recognition; SER）は、音声から人間の感情を識別する技術であり、メンタルヘルスケアやカスタマーサービス等において重要な役割を果たすため近年盛んに研究が行われている [1, 2, 3]。SER タスクは、カテゴリカルな分類とディメンショナルな分類に大別される [4]。カテゴリカルな感情分類は、発話を「喜び」「悲しみ」等の感情カテゴリに分類する、多クラス分類タスクである。一方、ディメンショナルな感情分類は、感情状態を感情価（Valence）、覚醒度（Arousal）、支配性（Dominance）といった複数の次元で表し、それぞれの数値を推定する回帰タスクである。本研究ではカテゴリカルな感情分類に焦点を当てる。

近年、多くの音声処理タスクにおいて、Transformer エンコーダ [5] に基づく事前学習済みの音声処理モデル（音声エンコーダ）の活用が検討されている。SER においても、wav2vec 2.0 [6] や HuBERT [7] などの自己教師あり学習（Self-supervised Learning; SSL）モデルに、感情認識結果を得るための小規模なデコーダを連結した手法が、従来手法を上回る精度を達成している [8, 9, 10]。wav2vec 2.0 などの SSL モデルは、大量のデータを用いた事前学習によって声の高さや音色などの**韻律**に関する豊富な知識を獲得しており、その知識が SER 等の下流タスクに役立てられていると考えられる。一方で、SSL モデルの持つ**言語**に関する知識は限定的であり、単語の意味や単語同士の関連（文脈）などの言語情報を十分考慮できない。そのため、SSL モデルに基づく SER は、韻律情報だけでは解くことが難しいケースの認識精度が低いことが知られている [10]。

音声エンコーダとして、大規模な音声認識（Automatic Speech Recognition; ASR）モデル Whisper [11] のエンコーダを用いた SER 手法も提案されており、SSL モデルに基づく手法と同等以上の精度を達成することが示されている [12, 13, 14, 15]。多言語 ASR や音声翻訳といった複数の音声言語処理タスクで教師あり学習された Whisper は、SSL モデルと比べて豊富な言語知識を持つと考えられる。またエンコーダ・デコーダ型の Transformer モデルである Whisper のデコーダは、単語の意味や文脈等を考慮してテキストを生成する、一種の言語モデルとしての機能を有する [16, 17]。そのため、Whisper デコーダを SER に利用することでより多くの言語情報を考慮した感情認識を行える可能性がある。しかし、Whisper に基づく既存手法 [12, 13, 14, 15] はいずれも Whisper エンコーダと小規模なデコーダで構成されており、Whisper デコーダの利用は検討されていない。

他のアプローチとして、韻律と言語情報の両方を考慮するため、音声とテキストを入力として用いるマルチモーダル SER[1)] が存在する [18, 19, 20, 21]。しかしこれらのモデルは、モダリティごとのエンコーダや外部の ASR モデルなど多くのモデルを結合するため構成が複雑であり、SER タスクで全体を最適

1）音声とテキストに基づく感情認識手法はマルチモーダル SER と呼ばれ、画像や生体信号等を含めた複数のモダリティを扱うマルチモーダル感情認識と区別される。



(a) 従来手法　　(b) 提案手法

図 1　Whisper に基づく SER。埋め込み層（Embedding layer）、位置埋め込み（Positional encoding）、出力の投射層（Projection layer）は省略されている。

化することが困難であるという課題があった。

本研究では、Whisper アーキテクチャ全体を活用した SER 手法である **Whisper-ER** を提案する。事前学習された Whisper デコーダの言語知識を利用することで、小規模なデコーダを用いた従来手法より多くの言語情報を考慮した感情認識が行えると考えた。提案手法は、Whisper の語彙を拡張し、感情クラスを表す追加トークン（例：“<|happy|>”）を導入する。その後、感情トークンを出力するようモデル全体を追加学習する。加えて、単一の系列で複数のタスクを学習するマルチタスク学習手法（SerialMTL）を提案する。ASR や性別認識（Gender Recognition; GR）などのサブタスクを学習させるマルチタスク学習（Multi-Task Learning; MTL）により、SER の性能が向上することが知られている [22, 23, 24, 23, 25]。SerialMTL では複数タスクのトークンを連結した単一の系列（例：“Really good . <|happy|>”）を推定する。従来の MTL を用いた手法と比べ、サブタスクごとにデコーダを追加する必要がなく、モデル構造がシンプルで拡張しやすい利点がある。また、従来のマルチモーダル SER と比べても、外部の ASR モデルやモダリティごとのエンコーダが不要であり、SER 精度に対して全体を最適化できる利点がある。

IEMOCAP [4] データセットを用いた実験で提案手法の有効性を検証した。Whisper-ER は従来手法の認識精度を相対的に 4.6%改善し、SerialMTL によって更に 6.3%の改善を得た。また、提案手法により、言語情報に強く依存する感情（“Happy”と “Angry”）の識別性能が大きく改善することを確認した。分析では、提案手法における Whisper デコーダが持つ事前学習済みの言語知識の重要性を検証した。

2　従来手法

本節では、Whisper エンコーダに基づく従来手法の概要を説明する（詳細は付録 A に示す）。従来手法のエンコーダは自己注意機構（Self-Attention）を備えた N 層のニューラルネットワークで構成される Whisper エンコーダであり、デコーダは平均プーリング層（Average pooling）と全結合層（FC layer）によって構成される（図 1(a)）。ここで、モデルの入力である音声特徴量を $\boldsymbol{X}$、対応する正解感情クラスを $c \in \{1, \cdots, C\}$ とする。C は感情クラスの総数である。従来手法のモデルは、エンコーダとデコーダを通して $\boldsymbol{X}$ から感情クラスの事後確率分布 $\boldsymbol{p} \in (0,1)^C$ を出力する。モデルは $\boldsymbol{p}$ と c 間の交差エントロピー誤差で学習され、推論時には $\boldsymbol{p}$ から最も確率の高いクラスを選択することで推定感情 $\hat{c}$ を得る。

3　提案手法

本節では、提案手法 Whisper-ER の概要を説明する（詳細は付録 B に示す）。Whisper-ER は Whisper アーキテクチャ全体を利用した SER モデルである。エンコーダには従来手法と同様に Whisper エンコーダを、デコーダには N 層のデコーダ層で構成される Whisper デコーダを使用する（図 1(b)）。各デコーダ層はエンコーダの出力をデコーダに取り込む相互注意機構（Cross-Attention）を備える。提案手法では Whisper の語彙を拡張して感情やサブタスクのクラスを表すトークンを追加する。例えば “<|happy|>” というトークンは感情クラス “Happy”を意味する。

本研究では 2 種類の学習方式、STL と SerialMTL を検討する。**STL** は、Whisper を単一のタスクで追加学習する学習手法であり、SER タスクの場合、モデルは感情ラベルに対応するトークンを出力するよう学習される。表 1 の 2 行目に STL で学習するトークン系列の一例を示す。**SerialMTL** はメインタスクとサブタスクの認識結果を含む単一のトークン系列を自己回帰的に生成するマルチタスク学習手法である。表 1 の 3 行目に SerialMTL で学習するトークン系列の一例を示す。例において、トークン系列には ASR (“Really good .”)、GR (“<|male|>”)、SER (“<|happy|>”) のトークン系列が含まれる。以降、このタスク設定を “ASR+GR+SER”と記述する。この設定では SER の出力が ASR や GR の出力に条件づけられるため、SerialMTL は言語情報（と性別情報）を考慮した一種のマルチモーダル SER とみなせる。

表 1　提案手法 Whisper-ER における正解トークン系列の例

	タスク	Token Sequence[a]
STL	SER	<|*sot*|> <|*en*|> <|*transcribe*|> <|*notimestamps*|> <|happy|> <|*eot*|>
SerialMTL	ASR+GR+SER	<|*sot*|> <|*en*|> <|*transcribe*|> <|*notimestamps*|> Really good . <|male|> <|happy|> <|*eot*|>

[a] トークン系列中の青字のテキストは目的タスクのトークン、黒いテキストは特殊トークンを示す。 最初の 4 トークンはプレフィックスとして与えられる。<|*sot*|> と <|*eot*|> は <|*startoftranscript*|> と <|*endoftext*|> の省略形。

4　実験

4.1　実験設定

実験では、SER の研究で広く用いられている IEMOCAP データセットを使用した [4, 8, 9, 23, 26]。IEMOCAP には 10 名のアメリカ英語話者による約 12 時間分の 1 対 1 会話の演技音声が含まれ、書き起こしおよび 3 名のアノテータによる感情ラベルが付与されている。先行研究 [27, 9] に従い、4 クラスのカテゴリカル感情分類（Neutral、Happy+Excited、Sad、Angry）を実施した。実験に使用した発話数は合計 5531（Neutral: 1708、Happy: 1636、Sad: 1084、Angry: 1103）である。学習データと評価データで話者が重複しない話者オープンの設定で実験を行い、1 組の話者ペアの会話を評価データ、1 組を検証データ、残り 3 組を学習データとして leave-one-out 交差検証を行った。

SER の手法として、従来手法（§2）と Whisper-ER の手法（STL、SerialMTL）（§3）を比較した。SerialMTL では、サブタスクが SER の結果に及ぼす影響を調べるため、ASR+SER、GR+SER、ASR+GR+SER の 3 通りを比較した。また、各手法において 2 つのサイズの Whisper（*base*[2] と *large-v3*[3]）を使用した。Whisper のエンコーダ及びデコーダ層の数は、*base* が 6 層、*large-v3* が 32 層である。学習可能なパラメータ数を付録 C に示す。学習は最大 24,000 ステップ行い、実効バッチサイズは 16 とした。480 ステップごとに精度検証を行い、モデルを保存した。早期終了は性能が 20 回連続で改善しない場合に適用した。最適化手法には AdamW（$\beta = 0.9$、$\beta_2 = 0.999$）を用い、学習率を $1e-5$ に設定の上、ReduceLROnPlateau スケジューラを利用した。MTL では SER の認識精度を精度検証に用い、STL ではそれぞれのタスクの認識精度を精度検証に用いた。学習終了後、検証に基づき上位 10 モデルを選び、それらのパラメータを混合して最終モデルを作成した。Whisper-ER の推論時には、ビーム幅 4 でビーム探索を行った。SER の評価指標として、発話ごとの精度の平均である Weighted Accuracy（WA）、およびクラスごとの精度の平均である Unweighted Accuracy（UA）を使用した。GR は Accuracy（Acc）、ASR では Word Error Rate（WER）で評価した。

2）https://huggingface.co/openai/whisper-base
3）https://huggingface.co/openai/whisper-large-v3

表 2　Whisper *base* と *large-v3* における SER の結果

Whisper			タスク	WA↑	UA↑
base	従来手法		SER	67.7	68.5
	Whisper-ER	STL	SER	71.9	72.7
		SerialMTL	A+S	72.4	72.9
			G+S	73.2	74.0
			A+G+S	**74.1**	**74.8**
large-v3	従来手法		SER	74.2	74.7
	Whisper-ER	STL	SER	77.3	78.1
		SerialMTL	A+S	78.2	78.9
			G+S	77.9	78.9
			A+G+S	**78.7**	**79.4**

4.2　実験結果

表 2 に各手法の SER の結果を示す。Whisper-ER STL は Whisper *base*、*large-v3* の両条件で従来手法を上回った。特に、Whisper *large-v3* の条件で、STL は UA 78.1%であり、従来手法を 4.6%相対的に改善した。また、Whisper-ER SerialMTL は両条件で STL を上回った。Whisper *large-v3* の条件で、SerialMTL は UA 79.4%であり、従来手法および STL に対してそれぞれ 6.3%と 1.7%の相対的な改善を得た。そのため、ASR と GR がそれぞれ SER の改善に寄与し、これらを組み合わせることでより大きな改善が得られたと考えられる。これらの結果は、先行研究とも整合している [22, 23]。以降は Whisper *large-v3* を用いた実験結果についてのみ述べる。

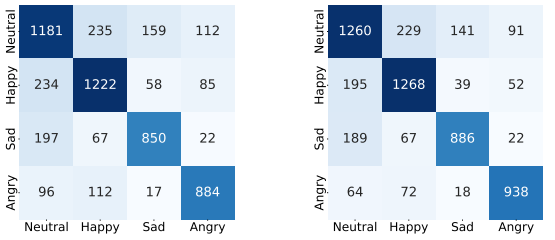
表 4 に Whisper-ER の STL（SER、GR、または ASR）と SerialMTL の比較を示す。上述の通り、SerialMTL による SER タスクの改善が見られた。SerialMTL の ASR 結果（WER 17.3%および 17.8%）は、追加学習を行う前の Whisper（3 行目、22.2%）と比べ改善しているものの、STL（16.7%）を下回った。GR に関

表 3　提案手法により SER 結果の改善が見られた例。

Audio (書き起こし)	正解	従来手法	Whisper-ER
(i) “We might be able to bring up a splendid little debate about this intemperate tots.”	Angry	Happy	Angry
(ii) “And it’s so close so you can come visit me all the time.”	Happy	Angry	Happy

表 4　SER とサブタスクの結果

		SER		GR	ASR
	タスク	WA↑	UA↑	Acc↑	WER↓
Whisper-ER	ASR	-	-	-	22.2
STL	SER	77.3	78.1	-	-
	GR	-	-	**99.6**	-
	ASR	-	-	-	**16.7**
SerialMTL	A+S	78.2	78.9	-	17.3
	G+S	77.9	78.9	99.4	-
	A+G+S	**78.7**	**79.4**	99.4	17.8



(a) 従来手法　(b) Whisper-ER (SerialMTL)

図 2　従来手法と提案手法における SER の混同行列

しては、STL と SerialMTL の間に有意な差は見られなかった。SerialMTL は単一モデルで複数の属性を推定でき、STL と近い精度を達成できる手法であると言える。そのため、感情や話者属性などの多くの情報を含んだ書き起こしを生成する簡便な手法として、今後の発展が期待される。

4.3　議論：言語情報を考慮できたか？

ディメンショナルな感情分類において、言語情報を考慮することで Valence の推定精度が向上することが知られている [10]。また感情を Valence と Arousal の二次元で表した感情の円環モデル [28] によれば、Happy と Angry は Valence の大小で区別することができる。従って、カテゴリカル感情分類においても、言語情報を考慮することで Happy と Angry の誤分類が減少すると予想される。図 2 は、従来手法と提案手法における混同行列を示したもので、提案手法は Happy と Angry の誤分類を従来手法に比べて 37.1%削減しており、全体の誤分類減少率 15.4%の中で大きい割合を占める。この結果は、提案手法が従来手法より多くの言語情報を考慮しており、それが精度改善に繋がったことを示唆している。

表 5　Whisper デコーダの事前学習の有無による精度変化の検証

	タスク	デコーダ	WA↑	UA↑
Whisper-ER	SER	事前学習あり	**77.3**	**78.1**
		事前学習なし	74.5	75.3
従来手法	SER	事前学習なし	74.2	74.7

表 3 は SER の認識結果の実例である。例 (i) の発話は “splendid”というポジティブな単語が含まれているが、文全体として皮肉や軽蔑を示しており、正解感情は Angry である。従来手法が誤って Happy を認識したことに対し、提案手法は正しい感情 Angry を認識した。また Happy と Angry の発話は韻律的特徴が似ており、区別するために言語情報が必要な場合が多い。(ii) は従来手法が Happy を Angry と誤って認識した例であり、この場合も提案手法は正しく認識できた。これらの結果から、提案手法はより多くの複雑な言語情報を考慮し、それが全体的な精度改善に繋がったと考えられる。

提案手法において Whisper デコーダの事前知識の有無が認識精度に与える影響を調査した。Whisper-ER STL において Whisper デコーダのパラメータをランダム初期化した場合、UA は 78.1%から 75.3%と従来手法と同等程度まで低下した（表 5）。そのため、Whisper デコーダの持つ言語等に関する事前知識が SER の推定に寄与したと考えられる。また、デコーダのサイズやアーキテクチャの違いが、従来手法と提案手法の性能差の主な要因ではないことも示している。

4.4　おわりに

本研究では音声認識モデル Whisper を用いた音声感情認識（SER）手法を提案した。提案手法は様々な音声言語処理タスクで学習された Whisper のデコーダを SER に利用することで、韻律情報と言語情報を考慮した感情認識を行う。加えて、SER と同時に音声認識や性別認識といったサブタスクを推定するマルチタスク学習手法を提案した。今後の展望として、サブタスクの性能低下を防ぐ手法の検討や、話者識別やディメンショナルな感情分類など様々なサブタスクを追加することなどが挙げられる。

参考文献

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. **IEEE Signal Processing Magazine**, Vol. 18, No. 1, pp. 32–80, 2001.

[2] R.W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: analysis of affective physiological state. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 23, No. 10, pp. 1175–1191, 2001.

[3] 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**, Vol. 28, pp. 715–728, 2020.

[4] 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, No. 4, pp. 335–359, 2008.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Advances in Neural Information Processing Systems**, Vol. 30, pp. 6000–6010, 2017.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 12449–12460, 2020.

[7] 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.

[8] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding. **arXiv preprint arXiv:2111.02735**, 2021.

[9] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. In **23th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 3400–3404, 2021.

[10] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 45, No. 9, pp. 10745–10759, 2023.

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. **arXiv preprint arXiv:2212.04356**, 2022.

[12] Tiantian Feng and Shrikanth Narayanan. PEFT-SER: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models. In **11th International Conference on Affective Computing and Intelligent Interaction (ACII)**, pp. 1–8, 2023.

[13] Mohamed Osman, Tamer Nadeem, and Ghada Khoriba. Towards generalizable SER: Soft labeling and data augmentation for modeling temporal emotion shifts in Large-Scale multilingual speech. **arXiv preprint arXiv:2311.08607**, 2023.

[14] Erik Goron, Lena Asai, Elias Rut, and Martin Dinov. Improving domain generalization in speech emotion recognition with whisper. In **2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 11631–11635, 2024.

[15] Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. EmoBox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In **26th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 1580–1584, 2024.

[16] Zhuo Gong, Daisuke Saito, Sheng Li, Hisashi Kawai, and Nobuaki Minematsu. Can we train a language model inside an end-to-end ASR model? - investigating effective implicit language modeling. In **Proceedings of the Second Workshop on When Creative AI Meets Conversational AI**, pp. 42–47, 2022.

[17] Yufei Liu, Rao Ma, Haihua Xu, Yi He, Zejun Ma, and Weibin Zhang. Internal language model estimation through explicit context vector learning for attention-based encoder-decoder ASR. In **24th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 1666–1670, 2022.

[18] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. Learning alignment for multimodal emotion recognition from speech. In **21th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 3569–3573, 2019.

[19] Yoonhyung Lee, Seunghyun Yoon, and Kyomin Jung. Multimodal speech emotion recognition using cross attention with aligned audio and text. In **22th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 2717–2721, 2020.

[20] Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition. In **22th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 3755–3759, 2020.

[21] Jarod Duret, Mickael Rouvier, and Yannick Estève. MSP-Podcast SER challenge 2024: L'antenne du ventoux multimodal Self-Supervised learning for speech emotion recognition. In **The Speaker and Language Recognition Workshop (Odyssey 2024)**, pp. 309–314, 2024.

[22] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church. Speech emotion recognition with multitask learning. In **23th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 4508–4512, 2021.

[23] Yuan Gao, Chenhui Chu, and Tatsuya Kawahara. Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with ASR and gender pretraining. In **25th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 3637–3641, 2023.

[24] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In **21th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 2803–2807, 2019.

[25] Yuan Gao, Hao Shi, Chenhui Chu, and Tatsuya Kawahara. Enhancing two-stage finetuning for speech emotion recognition using adapters. In **2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 11316–11320, 2024.

[26] Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. Vesper: A compact and effective pretrained model for speech emotion recognition. **IEEE Transactions on Affective Computing**, Vol. 15, No. 3, pp. 1711–1724, 2024.

[27] Aharon Satt, Shai Rozenberg, and Ron Hoory. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In **19th Annual Conference of the International Speech Communication Association (INTERSPEECH)**, pp. 1089–1093, 2017.

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

A　従来手法の詳細

従来モデル（§2）は、下式のように入力音声 $\boldsymbol{X}$ から感情クラスの事後確率分布 $\boldsymbol{p}$ を推定する。

$$\boldsymbol{H} = \mathrm{ENC}\left(\boldsymbol{X}; \boldsymbol{\theta}_e\right), \tag{1}$$

$$\boldsymbol{p} = \mathrm{softmax}\left(\mathrm{FC}\left(\mathrm{Pool}(\boldsymbol{H}); \boldsymbol{\theta}_f\right)\right), \tag{2}$$

$\boldsymbol{H} \in \mathbb{R}^{L \times D}$ はエンコーダによる長さ L の音声ベクトルであり、D はそのベクトル次元である。$\mathrm{ENC}(\cdot)$、$\mathrm{Pool}(\cdot)$、$\mathrm{FC}(\cdot)$ はそれぞれエンコーダ、平均プーリング、全結合層を表す。$\boldsymbol{\theta}_e$ と $\boldsymbol{\theta}_f$ はエンコーダとデコーダのパラメータ集合を表す。得られた事後確率 $\boldsymbol{p} = (p_1, p_2, ..., p_C)$ から最も確率の高いクラスを選択することで認識結果 $\hat{c}$ が得られる：$\hat{c} = \underset{i}{\mathrm{argmax}}(\boldsymbol{p})$。モデルは、$\boldsymbol{p}$ と c 間の交差エントロピー誤差 $\mathscr{L}$ の最小化を学習する。

$$\mathscr{L} = -\log P\left(c \mid \boldsymbol{X}\right), \tag{3}$$

$P(c|\boldsymbol{X}) = \boldsymbol{p}_c$ は $\boldsymbol{p}$ のうち正解クラス c に割り当てられた出力確率を示す。$\boldsymbol{\theta}_e$ の初期パラメータは事前学習済みの Whisper エンコーダから継承し、$\boldsymbol{\theta}_f$ はランダムな値を用いる。

B　提案手法の詳細

提案手法では、式 (1) のようにエンコーダで音声ベクトル $\boldsymbol{H}$ を抽出した後、出力トークン系列を自己回帰的に生成する。具体的には、i 番目の出力トークンに対する確率 $\boldsymbol{p}_i \in (0,1)^{V+W}$ は次式で計算される。

$$\boldsymbol{p}_i = \mathrm{softmax}(\mathrm{DEC}(\boldsymbol{H}, \hat{\boldsymbol{y}}_{1:i-1}; \boldsymbol{\theta}_d)), \tag{4}$$

V は Whisper の元の語彙数、W は追加トークンの数である。$\hat{\boldsymbol{y}}_{1:i-1}$ は過去に生成されたトークン列、$\mathrm{DEC}(\cdot)$ はデコーダ、$\boldsymbol{\theta}_d$ はデコーダのパラメータ集合を表す。$\boldsymbol{\theta}_d$ の初期パラメータは事前学習済みの Whisper デコーダから継承する。本研究では 2 種類の学習方式、STL と SerialMTL を提案する。

B.1　シングルタスク学習 (STL)

STL は、Whisper を単一の目的タスクで追加学習する学習手法である。SER タスクの場合、モデルは感情ラベルに対応するトークンを含む系列を出力するよう学習される（例：表 1 の 2 行目）。このトークン列には、言語タグやタイムスタンプタグなど、Whisper 特有の特殊トークンが含まれている。STL における損失関数 $\mathscr{L}_{\mathrm{STL}}$ は、ASR で用いられる交差エントロピー損失と同一である。正解のトークン列を $y = (y_1, y_2, \ldots, y_I)$, $\mathscr{L}_{\mathrm{STL}}$ とすると、$\mathscr{L}_{\mathrm{STL}}$ は以下で定義される。

$$\mathscr{L}_{\mathrm{STL}} = -\sum_{i=1}^{I} \log P(y_i \mid \boldsymbol{X}), \tag{5}$$

$P(y_i|\boldsymbol{X}) = \boldsymbol{p}_{i,y_i}$ は、$\boldsymbol{p}_i$ のうち正解トークン y_i に割り当てられた出力確率を示す。なお、上式において特殊トークンは <|*endoftext*|> を除き y に含めないため、SER タスクの場合 $I = 2$ となる。

B.2　マルチタスク系列学習 (SerialMTL)

SerialMTL ではメインタスクとサブタスクの認識結果を含む単一のトークン系列を自己回帰的に生成する。学習時の正解トークン系列は、$y = y^{(1)} \oplus y^{(2)} \oplus \cdots \oplus y^{(T)}$ のように定義される。$y^{(t)} = (y_1^{(t)}, \ldots, y_{I^{(t)}}^{(t)})$ $(t \in \{1, 2, \ldots, T\})$ は、t 番目のタスクに対応する長さ $I^{(t)}$ のトークン列である。$\oplus$ は左右の系列を連結する操作を表す。SerialMTL の損失関数 $\mathscr{L}_{\mathrm{SerialMTL}}$ は下式で定義される。

$$\mathscr{L}_{\mathrm{SerialMTL}} = \sum_{t=1}^{T} \left(-\frac{1}{I^{(t)}} \sum_{i=I^{(:t-1)}+1}^{I^{(:t)}} \log P(y_i \mid \boldsymbol{X}) \right) \tag{6}$$

$I^{(:t)} = \sum_{k=1}^{t} I^{(k)}$ は y に含まれる $y^{(t)}$ の終端のインデックスであり、$I^{(:0)} = 0$ とする。式 (5) との違いは、各タスクの損失を、そのタスクの系列長 $I^{(t)}$ で正規化してから合計する点である。これにより、ASR などの長い系列を持つタスクが支配的になることを防ぎ、タスクごとの系列長が異なっても、バランスの取れた学習が行える。

C　実験設定

表 6　各モデルの学習可能なパラメータ数

Whisper		パラメータ数
base	従来手法	20M
	Whisper-ER	72M
large-v3	従来手法	635M
	Whisper-ER	1.54B