博 士 論 文

多様な感情表現の理解に向けた
音声感情認識の高度化
（Advanced Speech Emotion Recognition for Understanding Various Emotion Expressions）

2025 年 3 月

立命館大学大学院 情報理工学研究科
情報理工学専攻 博士課程後期課程

永瀬　亮太郎

立命館大学審査博士論文

多様な感情表現の理解に向けた
音声感情認識の高度化
（Advanced Speech Emotion Recognition for Understanding Various Emotion Expressions）

2025 年 3 月
March 2025

立命館大学大学院 情報理工学研究科
情報理工学専攻 博士課程後期課程
Doctoral Program in Advanced
Information Science and Engineering
Graduate School of Information Science and Engineering
Ritsumeikan University

永瀬　亮太郎
NAGASE Ryotaro

研究指導教員：山下 洋一教授
Supervisor：Professor YAMASHITA Yoichi

概要

本論文では音声が伝える感情を推定する音声感情認識の研究に取り組む．この技術は，暮らしやビジネス，医療・福祉の幅広いサービスや商品に応用されている．

音声感情認識は多くの場合機械学習で実現される．特に，深層学習に基づく音声感情認識の研究は盛んに取り組まれており，認識率は年々向上している．これまでの研究では，音響情報のみを入力とし発話全体から喜怒哀楽などの代表的な感情を推定する手法が検討されてきた．そのため，音響情報のみでは誤認識しやすい感情や時間と共に変化する感情，代表的な感情のみでは示せない感情の認識が困難である．これらの感情を認識できるようにするため，本論文では「音響情報と言語情報を併用した音声感情認識」「感情ラベル列を用いた音声感情認識」「感情キャプションを活用した音声感情認識」の手法の検討を行った．

「音響情報と言語情報を併用した音声感情認識」の研究では，声高や周波数スペクトルなどの音響情報と発話した内容などの言語情報を組み合わせた手法を検討し，音声感情認識に効果的な手法を同一条件下で比較した．この実験を通して，認識性能の改善に効果的な音響・言語情報の組合せを明らかにした．また，言語情報として音声認識結果を利用し，2 つの情報の統合手法にどのような影響を与えるか明らかにした．

「感情ラベル列を用いた音声感情認識」の研究では，音素情報を利用し感情ラベルを並べた系列を予測するようにニューラルネットワークを学習する新しい手法を提案した．この手法によって，音素情報を活用することが細かな感情の認識に有効であることを明らかにした．また，発話内で感情が変化する音声を用意し，提案手法が時々刻々と変化する感情の認識に有効であることを示した．

「感情キャプションを活用した音声感情認識」の研究では，予測する感情形式として感情の説明文（感情キャプション）を利用し，音声から感情キャプションを推定する手法や推論時に分類するクラスを感情キャプションで自由に定義できる手法を提案した．この手法によって，代表的な感情では示せない感情をテキストとして認識できることを示した．また，感情に関連するクラス（e.g. 購買意欲の有無）などもある程度予測できることを示した．

以上の研究を通して，既存手法が認識していた感情の範囲を拡張した高度な音声感情認識の実現を目指した．

Abstract

This thesis tackles study of speech emotion recognition (SER), which predicts emotions conveyed by speech. SER has been applied to a wide range of services and products in daily life, business, healthcare, and welfare.

This technology is often implemented using machine learning methods. In particular, SER based on deep learning methods has been actively studied, leading to steady improvements in the accuracy of SER over the years. Previous studies have examined methods to predict basic emotions, such as happiness, anger, and sadness, from the whole utterance using only acoustic information as input. Therefore, it is difficult to recognize emotions that are often misrecognized using only acoustic information, emotions that change constantly, and emotions that cannot be represented by only basic emotions. To recognize these emotions, this thesis examined the methods of "SER using both acoustic and linguistic information," "SER with emotion label sequence," and "SER using emotion captions."

The study of "SER using both acoustic and linguistic information," investigated the methods combined acoustic information, such as pitch or spectrograms, and linguistic information, such as speech content. Moreover, these methods were compared under the same conditions. These experiments clarified the effective combination to improve the performance of SER. In addition, the experiments using the transcripts by automatic speech recognition as linguistic information clarified the effect of this on the combining methods.

The study of "SER with emotion label sequence" proposed new methods to train the neural network to predict a sequence of emotion labels using phoneme information. The results of these methods indicated that phoneme information is effective to recognize fine-grained emotions. In addition, the experiment using utterances in which the emotion changes demonstrated that the proposed methods are effective to recognize emotions that change constantly.

The study of "SER using emotion captions" used an emotion caption which describes emotions as a form of prediction, and proposed the method to predict emotion captions from speech or that to freely define classes by emotion captions during the predicting phase. The results showed that the model of predicting emotion captions can recognize the emotions which cannot be represented by basic emotions. Moreover, the experiment demonstrated that the model of predicting the class defined by emotion captions can predict classes related to emotions (e.g. purchase intention).

Through these studies, this thesis aims to realize the advanced SER which expands the range of the emotion recognized by previous methods.

目 次

図 目 次

表 目 次

第1章　序論

1.1　本研究の背景・目的

音声は情報伝達に必要な媒体である．音声が伝達する情報には発話内容などの言語情報や，感情などのパラ言語情報，発話した人の性別や年齢などの非言語情報が含まれる [1]．これらの情報を解析し，人と人及び人と機械のコミュニケーションを円滑にするための研究が盛んに取り組まれてきた．特に，パラ言語情報処理の一つである音声感情認識は HCI (human-computer interaction) からも注目されており，人と機械のコミュニケーションにとって重要な技術となっている．本研究では，この音声が伝える感情の認識についての研究に取り組む．この技術は既に様々なサービスや商品の研究開発に活用されている．例えば，人間に優しいロボットや対話エージェントの開発 [2,3]，コールセンタ音声の解析やコールセンタ支援システムの開発 [4,5]，e-learning 支援 [6]，メンタルヘルス分析 [7] などが挙げられる．このように音声感情認識は，日常生活やビジネス，教育，医療・福祉における音声のやり取りから人の感情状態を明らかにし，精神的により豊かな生活や社会の実現に必要な技術である．

音声感情認識の研究では，基本周波数や周波数スペクトルをはじめとする音響情報の取捨選択や機械学習手法の改善など数多くの手法が検討されてきた [8–10]．特に近年では，音響情報を含む時系列情報を扱う深層学習手法を活用した音声感情認識が盛んに検討されている [11–26]．これらの既存手法では，音響情報のみから分かる感情や発話単位の大まかな感情，喜怒哀楽などの代表的な感情を正確に認識することに注目されていた．そのため，音響情報のみでは誤認識しやすい感情や時間と共に変化する感情，代表的な感情のみでは示せない感情の認識は既存手法では困難である．図 1.1 には既存手法で認識できる感情本研究が目指す感情の範囲を示している．例えば，諦めの気持ちを込めて声高に「今日も居残りだ」と発話した音声から音響情報のみで感情認識した場合，結果が単に「喜び」となり「諦め」の感情が認識されない可能性がある．このような感情認識を対話システムが行った場合,「良かったですね」のような応答を行い，話者を不快にする可能性もある．また,「辛いから今すぐ休みたいけど，週末の旅行のために頑張るぞ」のように発話内で変化する感情音声から既存手法で感情認識した場合，結果が単に「悲しみ」となり，発話前半は「悲しみ」で発話後半は「期待」の感情が認識されない．このような感情認識を対話システムが行った場合,「それは残念ですね」のような応答を行い，話者がシステムに対して不審感を持つ可能性がある．他には，喜びや期待などを込めて「じゃあ出発しよう」と発話した音声から既存手法で感情認識した場合，結果が単に「喜び」となり，これから旅行に出発する高揚感や若干の不安感，前向きな気持ちなど具体的な感情は認識されない．このような感情認識を対話シ

ステムが行った場合,「いってらっしゃい」のような単純な応答を繰り返し，システムに対する話者の信頼感は生まれない可能性がある．人間に共感し信頼してもらえるコミュニケーションを機械で実現するには，これらの多様な感情を認識できるようにする必要がある．

故に，本論文では「音響情報と言語情報を併用した音声感情認識」「感情ラベル列を用いた音声感情認識」「感情キャプションを活用した音声感情認識」の手法の検討を行った．これらの研究を通して，図 1.1 の点線に示す通り既存手法が認識できる感情の範囲を拡張し，多様な感情を認識できる高度な音声感情認識の実現を目指す．

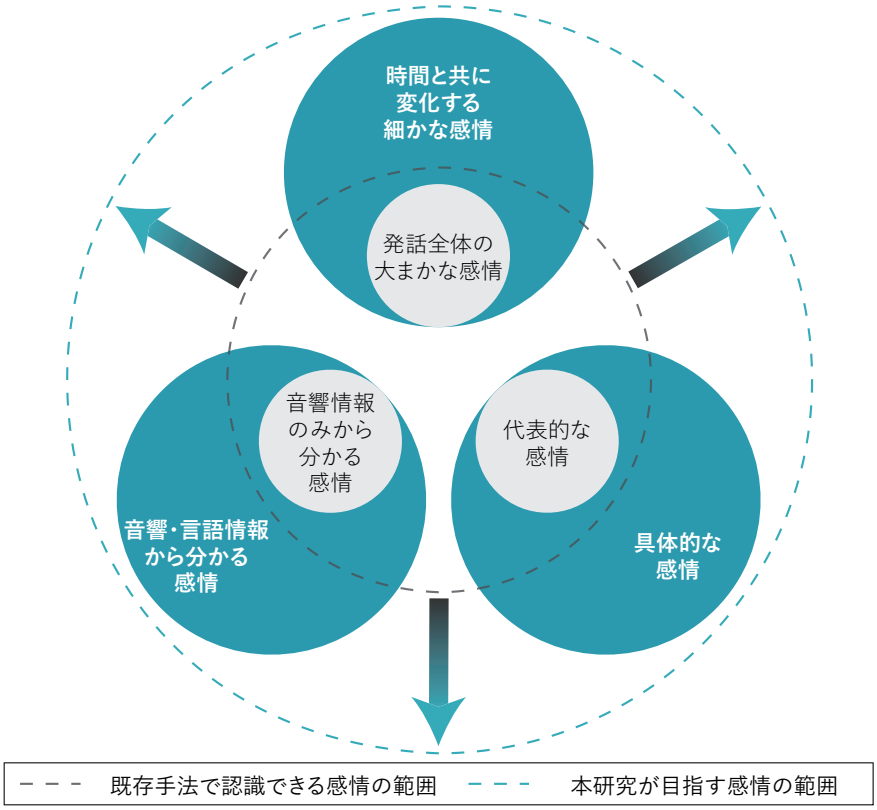


図 1.1: 本研究が目指す音声感情感情の高度化

1.2 本論文の構成

本論文は全 6 章で構成されている．第 2 章以降の構成を図 1.2 に示す．第 2 章では音声感情認識の定式化と一般的に用いられる評価指標について説明する．また，音声感情認識に用いられる主な深層学習技術とそれらを用いた関連研究について紹介する．最後に，第 3 章以降の研究の着眼点について説明する．第 3 章では音響情報と言語情報を併用した音声感情認識の研究について説明する．音響情報と言語情報の様々な統合処理を同一条件下で比較する．第 4 章では感情ラベルの系列を用いた音声感情認識の研究について説明する．発話に含まれる音素情報を活用した新しい手法を提案し，その効果を検証する．第 5 章では感情を説明した文を活用した音声感情認識の研究について説明する．代表的な感情だけでは示せない複雑な感情を推定する新しい手法を提案

し，その効果を検証する．第 6 章では本論文の結論を述べる．

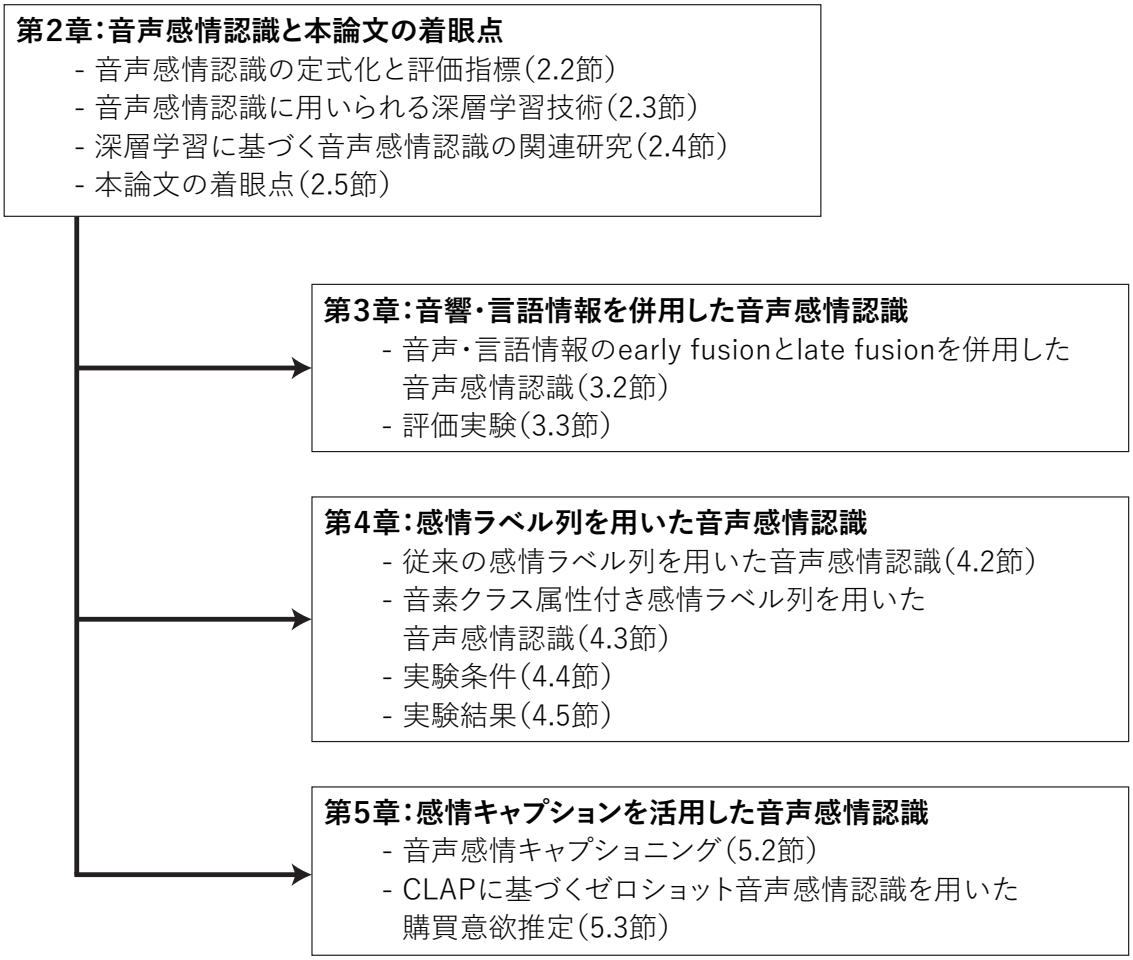


図 1.2: 本論文の構成

第2章　音声感情認識と本論文の着眼点

2.1　はじめに

本章では音声感情認識について，一般的な音声感情認識の定式化と評価指標について述べる．また，音声感情認識で用いられる深層学習技術と第 3 章以降の研究に関連した従来研究を説明する．最後に本論文の着眼点について述べる．

2.2　音声感情認識の定式化と評価指標

音声感情認識では主にカテゴリ感情または次元感情を推定することが多い [27]．カテゴリ感情とは「怒り」や「喜び」などのクラスで表現される感情 [28]，次元感情とは，ポジティブ－ネガティブなどの軸の値で表現される感情 [29] である．本研究では，音声からカテゴリ感情を推定する音声感情認識に取り組む．

2.2.1　問題の定式化

カテゴリ感情を推定する音声感情認識について，次の式 (2.1) に示す．ただし，$\mathbf{X}$ は入力特徴量，y は感情カテゴリ，$\hat{y}$ は予測した感情カテゴリとする．また，$P(y|\mathbf{X})$ は，任意の特徴量 $\mathbf{X}$ が入力されたときの，ある感情 y が認識される確率とする．

$$\hat{y} = \arg\max_{y} P(y|\mathbf{X}). \tag{2.1}$$

式 (2.1) では，$P(y|\mathbf{X})$ が最大になるときの y が $\hat{y}$ となる．カテゴリ感情を推定する音声感情認識では，この $P(y|\mathbf{X})$ をどれだけ正確に表現できるかによって，認識性能が決定する．$P(y|\mathbf{X})$ の予測には様々な統計的手法が用いられており，特に近年では深層学習手法が盛んに用いられている．

2.2.2　評価指標

音声からカテゴリ感情を認識する手法の評価では一般的な検出の評価指標が用いられる．検出における正解・予測と，評価指標の関係は，表 2.1 を用いて表される．この表を数値で表した行列のことを，混同行列 (confusion matrics) という．尚，TP は正解が真で予測も真と認識した頻度（真陽性），TN は正解が真で予測を偽と認識した頻度（真陰性），FP は正解が偽で予測を真と認

表 2.1: 分類における正解と予測の対応表

		Prediction Label	
		Positive (予測が真)	Negative （予測が偽）
True Label	Positive （正解が真）	TP	FN
	Negative （正解が偽）	FP	TN

識した頻度（偽陽性），FN は正解が偽で予測も偽と認識した頻度（偽陰性）を示す．これらの頻度から正解率や再現率，適合率，F 値が求められる．

正解率 (accuracy) は，全データに対して予測が正解した割合である．この評価指標は，音声感情認識の性能評価で最も用いられる．正解率の算出方法について，式 (2.2) に示す．

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}. \tag{2.2}$$

再現率 (recall) は，真に正解である場合のデータに対して予測が正解した割合である．再現率の算出方法について，式 (2.3) に示す．

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2.3}$$

適合率 (precision) は，正解と予測した場合のデータに対して実際に予測が正解した割合である．適合率の算出方法について，式 (2.4) に示す．

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.4}$$

F 値 (F-score) は，recall と precision の調和平均である．F 値の算出方法について，式 (2.5) に示す．ただし，n は重み係数であり，$n \in \mathbb{R}_0^+$ とする．音声感情認識の研究では，$n = 1$ のときの F 値 (F1-score) がよく用いられる．

$$\text{F}_\text{n}\text{-score} = (1 + n^2) \cdot \frac{\text{recall} \cdot \text{precision}}{n^2 \cdot \text{recall} + \text{precision}}. \tag{2.5}$$

2.3　音声感情認識に用いられる深層学習技術

2.3.1　CNN (convolutional neural network)

CNN (convolutional neural network) とは，入力情報にフィルタの重みを掛け合わせる層（畳み込み層）と条件に基づいて値を間引く層（プーリング層）によって構成されるネットワークである [30,31]．これは，D. H. Hubel と T. N. Wiesel の階層仮説 [32] に基づいて考案されたネットワークであり，畳み込み層とプーリング層は視覚領域のパターンの位置によって神経伝達が異なる 2 種類の細胞（単純型細胞と複雑型細胞）をモデル化している．この 2 種類の層を階層化することで視覚的な情報を取得できるネットワークを再現している．畳み込み層と代表的なプーリング層の一つである最大プーリング層の処理についてそれぞれ式 (2.6) と式 (2.7) に示し，CNN の概要を図 2.1 に示す．ただし，入力チャネル数を K，出力チャネル数を M，層のインデックスを l，カーネルのバイアスを b，プーリング結果を u とする．また，大きさ $I \times J$ の入力特徴量を $\mathbf{X}$，座標 (i,j) の入力特徴量を，$\mathbf{x}_{ij} = [x_{ij0}, \ldots, x_{ijk}, \ldots, x_{ijK}]$，大きさ $P \times Q$ の畳み込むフィルタ（カーネル）の重みを $\mathbf{W}^{(c)}$，座標 (p,q) のカーネルの重みを $\mathbf{w}_{pq}^{(c)}$，活性化関数を ϕ，入力特徴量の領域を $\mathcal{F}_{ij}$ とする．尚，出力チャネル数は用意したカーネルの数で決まる．故に m 番目のカーネル重みは $\mathbf{w}_{pqm}$，m 番目のカーネルから出力される座標 (i,j) の特徴量は x_{ijm} とする．

$$x_{ijm}^{(l)} = \phi\left(b_{ijm} + \sum_{p=0}^{P-1}\sum_{q=0}^{Q-1} \mathbf{x}_{(i+p)(j+q)}^{(l-1)} \cdot \mathbf{w}_{pqm}^{(c)}\right) \tag{2.6}$$

$$u_{ijk} = \max_{(p,q)\in\mathcal{F}_{ij}} x_{pqk}. \tag{2.7}$$

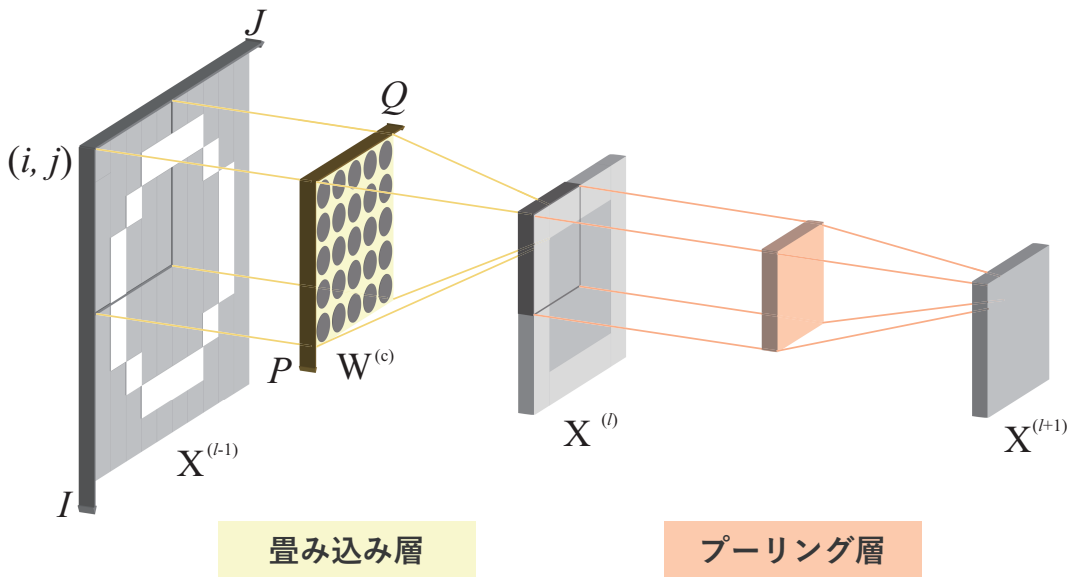
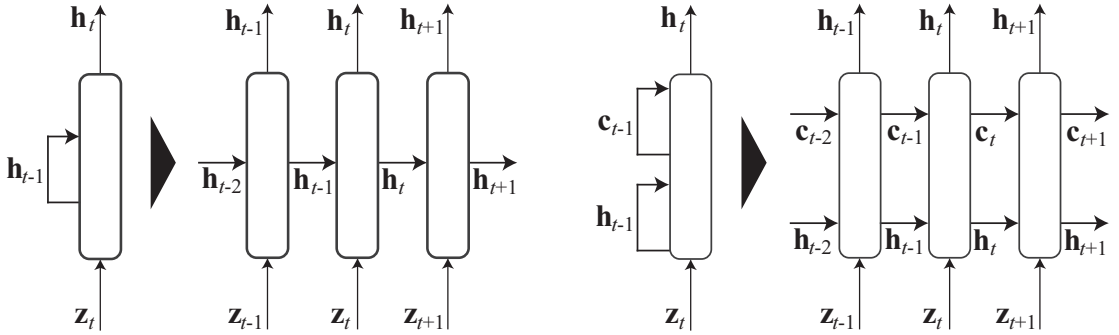


図 2.1: CNN の畳み込み層とプーリング層の概要図

2.3.2 RNN (recurrent neural network)

RNN (recurrent neural network) とは，一時刻前の内部状態（隠れ状態）を用いて現在の隠れ状態を算出するネットワークである [33]．再帰的な構造を持つネットワークの展開図を図 2.2 に示す．現在の時刻の隠れ状態は一時刻前の隠れ状態を用いて算出され，次の時刻に引き継がれる．



(a) RNN **(b)** LSTM

図 2.2: 再帰的な構造を持つネットワークの展開図

RNN の処理について式 (2.8) に示し，RNN の概要を図 2.3 に示す．ただし，入力長を T，時刻 t の入力信号を $\mathbf{Z} = [\mathbf{z}_0 \dots \mathbf{z}_t \dots \mathbf{z}_T]$，隠れ状態を $\mathbf{H} = [\mathbf{h}_0 \dots \mathbf{h}_t \dots \mathbf{h}_T]$，入力信号に対する重みを $\mathbf{W}^{(z)}$，隠れ状態に対する重みを $\mathbf{W}^{(h)}$，バイアスベクトルを $\mathbf{b}$ とする．

$$\mathbf{h}_t = \phi\left(\mathbf{W}^{(z)}\mathbf{z}_t + \mathbf{W}^{(h)}\mathbf{h}_{t-1} + \mathbf{b}\right) \tag{2.8}$$

RNN は過去の状態を参照できる一方，比較的長い時系列情報については勾配消失が起きやすいた

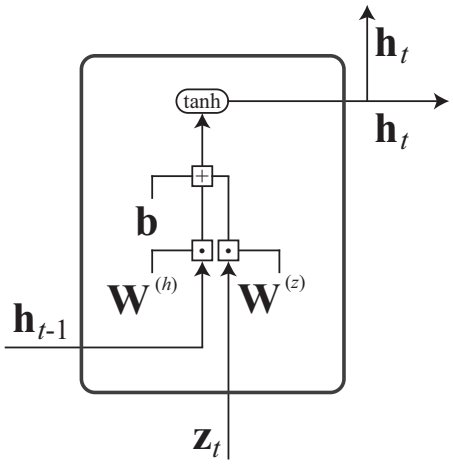


図 2.3: RNN の概要図

め，長期的な時系列の構造は捉えられないという問題があった．この課題を解決するために LSTM (long short term memory) が提案された．LSTM とは，勾配消失を低減するためにニューラルネットワークにゲート構造を取り入れた RNN の一つである [34]．ゲートには入力ゲートと出力ゲート，

忘却ゲートがあり，これらに併せて記憶セルをニューラルネットワークに導入することで，RNN よりも長期間の情報を保持しながら学習を進めることができる．LSTM の記憶セル，入力ゲート，出力ゲート，忘却ゲートの処理をそれぞれ式 (2.9)〜式 (2.12) に，出力される記憶セルと隠れ状態の更新式をそれぞれ式 (2.13) と式 (2.14) に示し，LSTM の概要を図 2.4 に示す．ただし，新しい記憶の生成結果を $\mathbf{r}_t$，入力ゲートの出力結果を $\mathbf{i}_t$，出力ゲートの出力結果を $\mathbf{o}_t$，忘却ゲートの出力結果を $\mathbf{f}_t$，更新された記憶セルを $\mathbf{c}_t$ とする．また，入力特徴量に対する学習可能な重みを $\mathbf{W}^{(z,r)}$，$\mathbf{W}^{(z,i)}$，$\mathbf{W}^{(z,o)}$，$\mathbf{W}^{(z,f)}$，隠れ状態に対する学習可能な重みを $\mathbf{W}^{(h,r)}$，$\mathbf{W}^{(h,i)}$，$\mathbf{W}^{(h,o)}$，$\mathbf{W}^{(h,f)}$，各処理についてのバイアスベクトルを $\mathbf{b}^{(r)}$, $\mathbf{b}^{(i)}$, $\mathbf{b}^{(o)}$, $\mathbf{b}^{(f)}$ とする．$\circ$ はアダマール積とする．

$$\mathbf{r}_t = \phi\left(\mathbf{W}^{(z,r)}\mathbf{z}_t + \mathbf{W}^{(h,r)}\mathbf{h}_{t-1} + \mathbf{b}^{(r)}\right) \tag{2.9}$$

$$\mathbf{i}_t = \phi\left(\mathbf{W}^{(z,i)}\mathbf{z}_t + \mathbf{W}^{(h,i)}\mathbf{h}_{t-1} + \mathbf{b}^{(i)}\right) \tag{2.10}$$

$$\mathbf{o}_t = \phi\left(\mathbf{W}^{(z,o)}\mathbf{z}_t + \mathbf{W}^{(h,o)}\mathbf{h}_{t-1} + \mathbf{b}^{(o)}\right) \tag{2.11}$$

$$\mathbf{f}_t = \phi\left(\mathbf{W}^{(z,f)}\mathbf{z}_t + \mathbf{W}^{(h,f)}\mathbf{h}_{t-1} + \mathbf{b}^{(f)}\right) \tag{2.12}$$

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{r}_t \circ \mathbf{i}_t \tag{2.13}$$

$$\mathbf{h}_t = \mathbf{o} \circ \phi(\mathbf{c}_t) \tag{2.14}$$

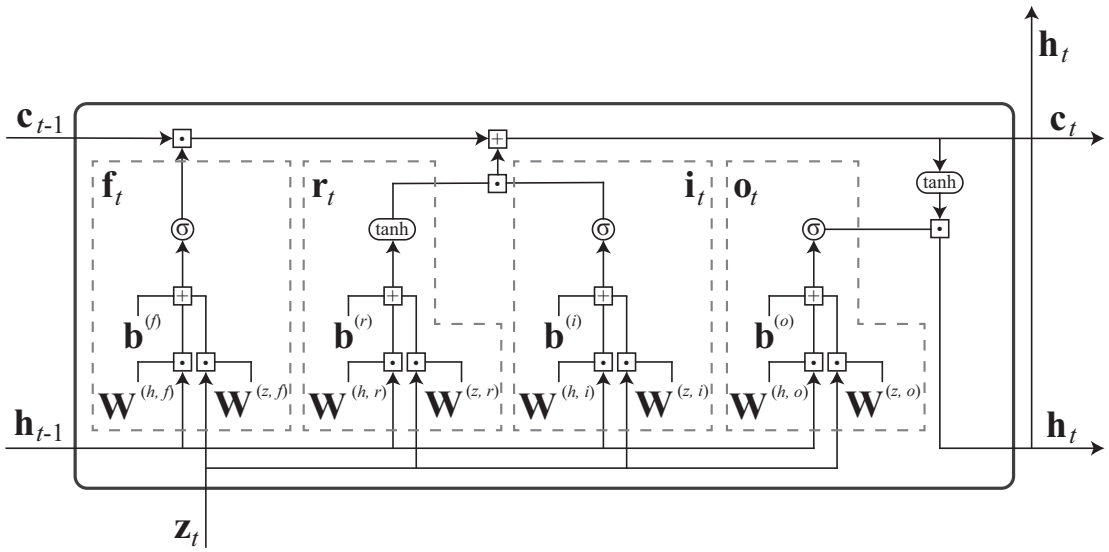


図 2.4: LSTM の概要図

また，順方向と逆方向の LSTM を組み合わせ双方向の隠れ状態を参照する BLSTM (bidirectional LSTM) も提案されている．現在の隠れ状態を前後の隠れ状態を用いて算出することで過去と未来の時間依存性を捉えることが可能になっている．

2.3.3 注意機構

注意機構 (attention mechanism) [35, 36] とは，入力情報の内，注目すべき特徴を算出するネットワークである．特に，入力が同じ時系列情報になる注意機構を自己注意機構 (self attention mechanism) という．注意機構は式 (2.15) の通りである．ただし，注意機構の出力を $\mathbf{S}$，注意重み (attention weight) を $\mathbf{A}$ とする．

$$\mathbf{S} = \mathbf{A}\mathbf{H}^{\mathrm{T}} \tag{2.15}$$

注意重みの算出方法には様々なものがある．本節では代表的な手法として Zhouhan らの手法 [35] と Vaswani らの手法 [36] を取り上げる．

Zhouhan らは，文から解釈可能な固定長の埋め込み表現を得るための自己注意機構を提案した [35]．Zhouhan らの自己注意機構における注意重みの算出方法は式 (2.16) に示す通りである．ただし，隠れ状態に対する学習可能な重みを $\mathbf{W}^{(a1)}$，$\mathbf{W}^{(a2)}$ とする．

$$\mathbf{A} = \mathrm{softmax}(\mathbf{W}^{(a2)} \tanh(\mathbf{W}^{(a1)}\mathbf{H})) \tag{2.16}$$

Zhouhan らはこの自己注意機構を用いて著者プロファイリングや感情分析，テキスト含意の推定モデルを構築し，従来手法と同等以上の性能を達成した．

また，Vaswani らは機械翻訳の性能向上を目的に Transformer [36] を提案し，その中で勾配消失を軽減した注意機構である scaled dot-product attention を利用されている．尚，Transformer とは，CNN や RNN を利用せず埋め込み表現の位置情報を示す positional encoder と注意機構のみで時系列情報を扱うことが出来る encoder-decoder モデルである．Vaswani らの注意重みの算出方法は式 (2.17) に示す通りである．ただし，$\mathbf{H}$ の次元数を d_h とする．

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{H}^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}})^{\mathrm{T}}}{\sqrt{d_h}}\right) \tag{2.17}$$

2.4 深層学習に基づく音声感情認識の関連研究

2.4.1 音声感情認識の入力特徴量

音声感情認識の入力には，音声から分析した様々な特徴量を利用する．例えば，声の高さを表現した基本周波数や音声を短時間フーリエ変換することで得られる周波数スペクトログラム，人間の聴覚特性を考慮した尺度（メル尺度）で調音の周波数特性を表現したメル周波数ケプストラム係数などが用いられる．また，音声から分析して得られる特徴の内，感情表現と関わりのあるものを選択した特徴量セットである低次記述子 (IS09 [8]，ComParE2016 [9]，eGeMAPS [10]) なども利用される．近年では，音声波形自体や音声波形をニューラルネットワークに入力して得られる深層埋め込みなどが音声感情認識の入力として利用される．

2.4.2 発話単位の音声感情認識

音声感情認識の研究ではこれまでに数多くの手法が検討されてきた．特に近年，深層学習を用いた音声感情認識の研究が盛んに取り組まれている．例えば，Badshah らは階層型 CNN を用いてスペクトログラムから感情を明確に識別できる特徴を捉えるための手法を提案し，認識率の向上を目指した [11]．また，Wöllmer らは BLSTM を用いた手法を提案し，時間と共に変化する感情に関する情報を組み込んだニューラルネットワークによって認識率の向上を目指した [12]．

Transformer が提案されて以降は，CNN 及び RNN と併せて注意機構も用いられるようになった．例えば，Xie らは注意機構と LSTM を用いた手法を提案し，LSTM のみを用いた手法よりも高い認識率を達成している [13]．また，Li らは CNN と BLSTM，自己注意機構で構成された認識器を性別分類と感情認識のマルチタスクで学習する手法を提案している [14]．この他にも，様々な入力特徴量や CNN，BLSTM，注意機構を用いたモデル構造などが提案されている [15–17]．

wav2vec2.0 や HuBERT などをはじめとする自己教師あり学習に基づく音声埋め込み表現モデルが提案されて以降は，事前学習済み自己教師あり学習モデルを活用した研究が取り組まれるようになった．例えば，Pepino らは事前学習済み自己教師あり学習モデルから得られた音声埋め込み表現を入力とする BLSTM を用いた手法を提案し，低次記述子やスペクトログラムを用いた場合よりも正確な認識率を達成した [18]．また，Cai らは自己教師あり学習モデルを利用した音声認識と音声感情認識のマルチタスク学習手法を提案し，感情の認識性能を改善した [19]．この他にも，多くの研究で事前学習済み自己教師あり学習モデルが用いられており，音声感情認識の性能改善が確認されている [20–24]．

2.4.3 フレーム単位の音声感情認識

音声感情認識では多くの場合，入力された発話に対して一つの感情ラベルを推定するように認識器を学習している．この場合，入力発話の全区間に対して均一に正解感情を与えているため，本来感情が表出していない区間や発話内で時間と共に変化する感情状態の認識は困難である．そのため，従来研究では入力発話に対して正解感情の系列（感情ラベル列）を用いて認識器を学習する手法が提案されている．例えば，Fayek らは入力音声の発話区間には感情ラベルを，そうでない区間には無音ラベルを付与した感情ラベル列を用いて認識器を学習する手法を提案した [25]．また，Han らは入力音声内の有声音素の数だけ並べた感情ラベル列を入力音声から推定するニューラルネットワークの学習方法を提案した [26]．入力発話から感情ラベル列を推定する手法では感情状態と無感情状態を数フレーム毎に推定するため，発話内で時間と共に変化する感情状態の認識が比較的容易である．尚，フレーム単位の音声感情認識の研究は取り組まれてきたが，これらは全て発話単位で評価されており，フレーム単位の感情の認識性能は十分評価されていない．

2.5　本論文の着眼点

第 2.4 節で紹介した通り，音声感情認識は様々な改良を経て認識率が年々向上している．しかしながら，既存手法の認識率の向上だけでは認識が困難な感情が存在する．

例えば，音響情報のみでは誤認識するような感情は依然として認識が困難である．多くの研究で音響情報のみを入力とした音声感情認識が検討されている．そのため,「惜しかったね」と言っているが悲しんで聞こえないような音声から感情を推定しようとすると，悲しみではない感情として分類される場合がある．

他には，時々刻々と変化するような感情の認識も依然として困難である．感情が変化する音声からフレーム単位で正しく感情を推定できるかは十分評価されていない．そのため,「待ちに待った休日が他の予定で潰れた」のように前半の期待感のあるポジティブな感情と後半の落ち込んでいるネガティブな感情をフレーム単位で分類できるのか明らかになっていない．

また，代表的な感情のみでは表現できない感情も認識が困難である．多くの研究では，カテゴリ感情や次元感情を推定するように認識器を構築する．そのため,「勝利に興奮し満足感を感じている」のような具体的な感情を認識することは困難である．

本研究ではこれらの 3 つの感情の認識を目指し，新たな音声感情認識手法を検討した．

第3章　音響・言語情報を併用した音声感情認識

3.1　はじめに

近年，音響情報だけでなく単語や文の意味などを示す言語情報も併用した音声感情認識の研究が盛んに取り組まれている．音響情報のみを用いた場合と言語情報のみを用いた場合の感情認識の認識傾向については，従来研究で示されている [37]．音響情報のみを用いた 4 感情（怒り，喜び，悲しみ，平静）の認識の場合，平静の認識率は高いが喜びは平静に誤分類されやすい．一方で，言語情報のみを用いた 4 感情の認識の場合，喜びと平静の発話に含まれる単語の違いを捉えられるため喜びの認識率が向上する．故に，音響情報と言語情報を併用することでそれぞれの利点を生かし，認識率の向上が可能であるとされている．

音響情報と言語情報のような 2 つの異なる情報の統合には次の 2 つの手法が考えられる．1 つ目は特徴量の融合 (early fusion) である．これは発話全体から得られた低次な特徴量同士を統合することを指す．2 つ目は予測結果の融合 (late fusion) である．これは発話全体から得られた高次な予測結果同士を統合することを指す．従来研究では，early fusion または late fusion を利用して音声感情認識に言語情報を取り入れている．例えば Atmaja らは音声とテキストから得られた埋め込み表現を early fusion する手法を提案した [38]．Cho らは音響情報を用いた感情認識と言語情報を用いた感情認識から得られた予測結果を late fusion し，英語音声からの感情の認識性能を向上させた [39]．

Early fusion を利用する場合，音響・言語情報の特徴量間の関係を利用して感情を決定することができる．しかし，音響・言語情報のいずれかで感情表出が不明瞭な場合，融合後の特徴量から判断できる感情も不明瞭になるため，様々な音声が平静に偏って分類される可能性がある[1]．一方 late fusion を利用する場合，音響・言語情報それぞれで予測した結果を融合するため，いずれかの情報で感情表現が明瞭であれば感情を正しく決定できると考えられる．例えば,「悲しみ」の感情で発話された「よろしくお願いします」を考えると，late fusion を用いることで，言語情報における感情表出が明確でなくても音響情報における感情表出が明らかであれば正しく感情を決定できる可能性がある．故に，これらの fusion を併用することで音響・言語情報から認識できる感情音声が増加し全体的な正解率が向上すると期待できる．これまでにも early fusion と late fusion を併用した手法が提案されてきた．例えば Chen らは early fusion に特徴量の結合，late fusion に予測結果の和を用いて感情を認識する手法 [40]，Pepino らは early fusion に特徴量の結合，late

[1] 予備実験において early fusion のみを用いた場合では，喜びや怒り，悲しみなどの感情音声を平静に誤分類する傾向がみられた．

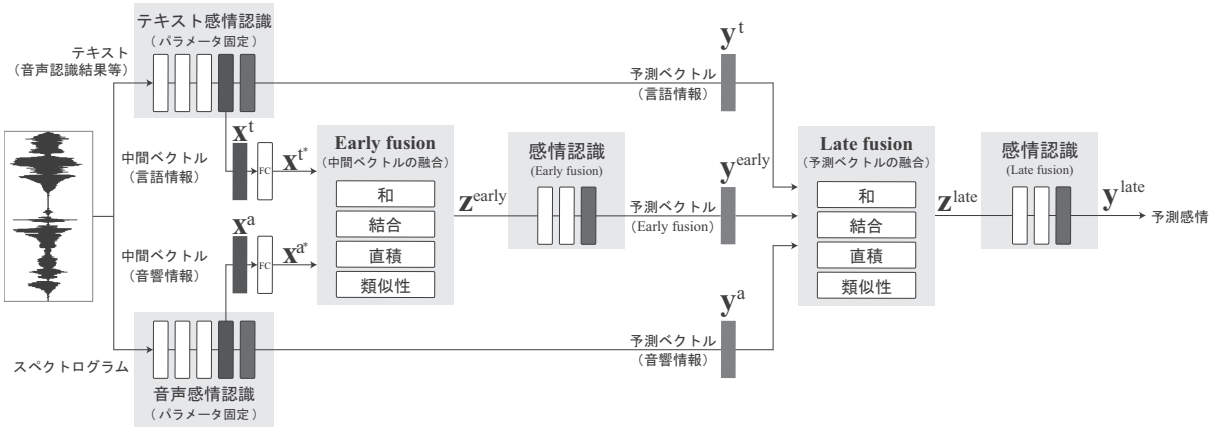


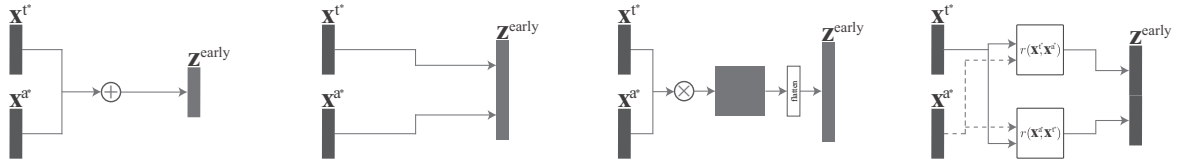
図 3.1: Early fusion と late fusion を用いた音声感情認識器の概要

fusion に予測結果の結合を用いて感情を認識する手法 [41] を提案している．しかし上記の研究ではこれらの組み合わせ以外は比較，検討されていない．Early fusion のみ，または late fusion のみを用いた研究では様々な融合手法を比較し，その有効性を検証しているため，early fusion と late fusion を併用した手法も様々な融合手法と比較する必要がある．また，同じデータセットや学習，検証，評価手法で音声感情認識およびテキスト感情認識から得られる情報の様々な融合の処理を併用した手法を比較した実験もされていない．特に，話者だけでなく発話も完全に重複なく，全データを学習，検証，評価に利用した実験はされていない．故に本研究では，音響情報に基づく音声感情認識とテキスト感情認識から得られた情報を early fusion と late fusion を併用し融合する新しい手法を提案する．また，4 種類の異なる融合処理を early fusion と late fusion に利用し，同一条件下で評価，比較を行なう．

3.2 音響・言語情報の early fusion と late fusion を併用した音声感情認識

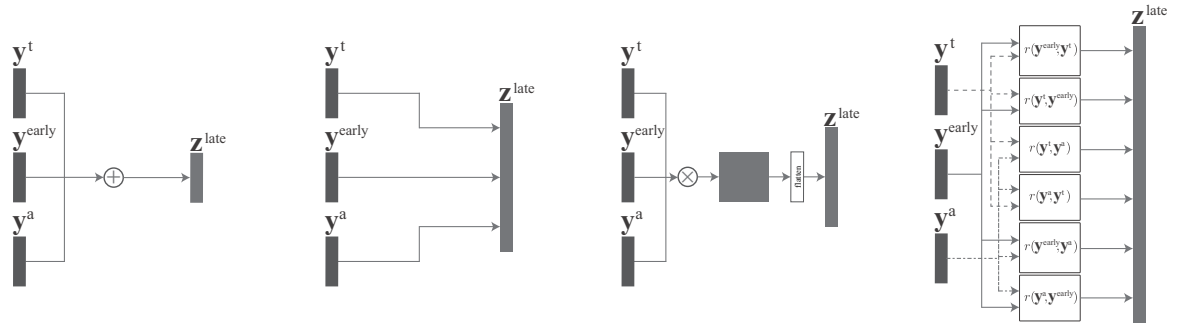
本章では，本研究で用いた early fusion と late fusion を併用した手法について説明する．表 3.1 は融合手法の一覧，図 3.1 は融合手法を用いた音声感情認識器の概要を示している．尚，表 3.1 の No. は各手法の ID 番号，[Ours] は提案手法，白丸と黒丸はそれぞれ融合するベクトルが 2 種類または 3 種類の場合を示す．また，従来研究の一部の処理を利用した融合手法については No. の横に該当の文献を付記している．

本論文において early fusion とは発話全体から得られた異なる性質を持つ 2 つ以上の特徴量を融合することである．この融合により，特徴量同士の関係性を考慮して認識器を学習できるようになる．一方，late fusion とは発話全体から得られた異なる性質を持つ 2 つ以上の予測結果を融合することである．入力情報に依らず同じ形式で出力される予測結果を融合するため，異なる形式を持つ特徴量を対象とした early fusion に比べ融合が容易である．本研究では，early fusion と late fusion を組み合わせ，特徴量同士または予測結果同士の関係性を考慮することで更なる認識



(a) 和による融合　**(b)** 結合による融合　**(c)** 直積による融合　**(d)** 類似性に基づいた融合

図 3.2: 各融合処理の概要 (early fusion)



(a) 和による融合　**(b)** 結合による融合　**(c)** 直積による融合　**(d)** 類似性に基づいた融合

図 3.3: 各融合処理の概要 (late fusion)

性能の向上を目指す．Early fusion または late fusion には，深層学習に基づく音声感情認識とテキスト感情認識の事前学習済み認識モデルを利用した．融合する情報は，各認識モデルの最終層の一つ手前の層から得られるベクトル（中間ベクトル）と最終層から得られるベクトル（予測ベクトル）とする．

3.2.1　予備的検討

本節では early fusion と late fusion の併用による認識性能向上の可能性を調査するために，early fusion と late fusion を個別に用いた音声感情認識の実験を実施した．Early fusion では音響情報を用いた音声感情認識モデルと言語情報を用いたテキスト感情認識モデルの中間ベクトル同士を，late fusion では予測ベクトル同士を結合した．尚，融合処理にはそれぞれ No. 2 と No. 22 を用いた．得られた認識結果における正解/不正解のサンプル数の関係を表 3.2 に示す．結果より，一方の fusion では正解し，他方の fusion では不正解であったサンプル数は合計で 6,356 であり，全体の約 31.8%であることがわかった．これは early fusion の対象である中間ベクトルと late fusion の対象である予測ベクトルは包含関係にないことを示している．また，early fusion と late fusion を併用することで正しく認識できるサンプル数が増加する可能性を示している．

表 3.1: 融合手法の一覧

No.	Early fusion				Late fusion			
	I	II	III	IV	I	II	III	IV
1	○							
2 [20]		○						
3 [42]			○					
4 [23]				○				
5 [Ours]	○				●			
6 [Ours]	○					●		
7 [Ours]	○						●	
8 [Ours]	○							●
9 [40]		○			●			
10 [41]		○				●		
11 [Ours]		○					●	
12 [Ours]		○						●
13 [Ours]			○		●			
14 [Ours]			○			●		
15 [Ours]			○				●	
16 [Ours]			○					●
17 [Ours]				○	●			
18 [Ours]				○		●		
19 [Ours]				○			●	
20 [Ours]				○				●
21 [43]					○			
22 [39]						○		
23							○	
24								○

※ I：和，II：結合，III：直積，IV：類似性

表 3.2: Early fusion と late fusion を個別に用いた場合の認識結果における正解/不正解数の関係

		Late fusion	
		正解	不正解
Early fusion	正解	10,142	3,760
	不正解	2,596	3,502

3.2.2 提案手法

本節では，early fusion または late fusion で用いる 4 種類の融合処理について説明する．図 3.2, 3.3 は early fusion 及び late fusion での各融合手法の概要を示している．尚，$\mathbf{x}^{\mathrm{a}} = [x_1^{\mathrm{a}}, \ldots, x_{N_{\mathrm{a}}}^{\mathrm{a}}]^\top$ と $\mathbf{x}^{\mathrm{t}} = [x_1^{\mathrm{t}}, \ldots, x_{N_{\mathrm{t}}}^{\mathrm{t}}]^\top$ はそれぞれ音響情報と言語情報から得られる中間ベクトル，$\mathbf{x}^{\mathrm{a}^*} = [x_1^{\mathrm{a}^*}, \ldots, x_{N_{\mathrm{a}^*}}^{\mathrm{a}^*}]^\top$ と $\mathbf{x}^{\mathrm{t}^*} = [x_1^{\mathrm{t}^*}, \ldots, x_{N_{\mathrm{t}^*}}^{\mathrm{t}^*}]^\top$ は各情報の中間ベクトルを全結合層に入力し得られたベクトル，$\mathbf{y}^{\mathrm{a}} = [y_1^{\mathrm{a}}, \ldots, y_M^{\mathrm{a}}]^\top$ と $\mathbf{y}^{\mathrm{early}} = [y_1^{\mathrm{early}}, \ldots, y_M^{\mathrm{early}}]^\top$，$\mathbf{y}^{\mathrm{t}} = [y_1^{\mathrm{t}}, \ldots, y_M^{\mathrm{t}}]^\top$，$\mathbf{y}^{\mathrm{late}} = [y_1^{\mathrm{late}}, \ldots, y_M^{\mathrm{late}}]^\top$ はそれぞれ音響情報，early fusion，言語情報，late fusion から得られる予測ベクトルとする．各情報の中間ベクトル $\mathbf{x}^a$，$\mathbf{x}^t$ はそれぞれ次元数が異なるため，early fusion する前に全結合層を適用し次元数を揃えている．一方，各情報の予測ベクトルは次元数が揃っているため，そのまま late fusion に利用している．また，N_{a}，N_{t} はそれぞれ音響情報と言語情報から得られる中間ベクトルの次元数，N_{a^*}，N_{t^*} は各情報の中間ベクトルを全結合層に入力し得られたベクトルの次元数，M は予測ベクトルの次元数とする．Early fusion と late fusion を併用した場合の late fusion の式は，式 (3.2) と式 (3.4)，式 (3.6)，式 (3.9) に示す．ただし，Late fusion のみの場合は $\mathbf{y}^{\mathrm{early}}$ を除く 2 種類の予測ベクトルの融合となる．

1. **和による融合**：和による融合では図 3.2a，3.3a に示す通り，同じ大きさのベクトルを要素ごとに足し合わせる．最も単純な融合であり，形式が同じベクトル同士の融合に有効である．従来研究では late fusion に用いられている [43]．和による融合を式 (3.1) と式 (3.2) に示す．ただし，$\mathbf{z}_{\mathrm{sum}}^{\mathrm{early}}$ は early fusion 後のベクトル，$\mathbf{z}_{\mathrm{sum}}^{\mathrm{late}}$ は late fusion 後のベクトルとする．

$$\mathbf{z}_{\mathrm{sum}}^{\mathrm{early}} = \mathbf{x}^{\mathrm{a}^*} + \mathbf{x}^{\mathrm{t}^*} \tag{3.1}$$

$$\mathbf{z}_{\mathrm{sum}}^{\mathrm{late}} = \mathbf{y}^{\mathrm{a}} + \mathbf{y}^{\mathrm{early}} + \mathbf{y}^{\mathrm{t}} \tag{3.2}$$

2. **結合による融合**：結合による融合では図 3.2b，3.3b に示す通り，ベクトルを連結する．この融合により，ベクトルの値に重みを付けてモデルを最適化できる．従来研究では入力特徴量や事前学習済み感情認識器から得られる埋め込み表現の early fusion に用いられている [20, 41]．結合による融合を式 (3.3) と式 (3.4) に示す．ただし，$\mathbf{z}_{\mathrm{cat}}^{\mathrm{early}}$ は early fusion 後のベクトル，$\mathbf{z}_{\mathrm{cat}}^{\mathrm{late}}$ は late fusion 後のベクトルとする．

$$\mathbf{z}_{\mathrm{cat}}^{\mathrm{early}} = [\mathbf{x}^{\mathrm{a}^{*\top}}, \mathbf{x}^{\mathrm{t}^{*\top}}] \tag{3.3}$$

$$\mathbf{z}_{\mathrm{cat}}^{\mathrm{late}} = [\mathbf{y}^{\mathrm{a}\top}, \mathbf{y}^{\mathrm{early}\top}, \mathbf{y}^{\mathrm{t}\top}] \tag{3.4}$$

3. **直積による融合**：直積による融合では図 3.2c，3.3c に示す通り，各ベクトルの直積を算出し平坦化する．単なる重み付け和ではなく各ベクトルの直積を算出することで，入力情報の関連性に着目してモデルを最適化できる．従来研究では音響情報や言語情報，視覚情報など複数の特徴量を入力する early fusion に用いられている [42]．直積による融合を式 (3.5) と式 (3.6) に示す．ただし，flatten は平坦化層，$\otimes$ は直積，$\mathbf{z}_{\mathrm{tensor}}^{\mathrm{early}}$ は early fusion 後のベクトル，

$\mathbf{z}_{\text{tensor}}^{\text{late}}$ は late fusion 後のベクトルとする．

$$\mathbf{z}_{\text{tensor}}^{\text{early}} = \text{flatten}([\mathbf{x}^{\text{a}^*} \otimes \mathbf{x}^{\text{t}^*}]) \tag{3.5}$$

$$\mathbf{z}_{\text{tensor}}^{\text{late}} = \text{flatten}([\mathbf{y}^{\text{a}} \otimes \mathbf{y}^{\text{early}} \otimes \mathbf{y}^{\text{t}}]) \tag{3.6}$$

4. **類似性に基づいた融合**：類似性に基づいた融合では図 3.2d，3.3d に示す通り，入力ベクトル間の類似性を示す値を算出する．これにより異なる入力情報の関係性を明示的に最適化できる．関連研究として Attention を用いた融合を early fusion に適用した事例がある [23]．類似性を示す値は式 (3.7) で算出される．類似性を示す値の算出過程は図 3.4 に示す通りである．ただし，fc は一層の全結合層，softmax は行方向のソフトマックス関数，$\mathbf{v} = [v_1, \ldots, v_L]^\top$，$\mathbf{w} = [w_1, \ldots, w_L]^\top$ はそれぞれ次元数 L のベクトルとする．

$$r(\mathbf{v}, \mathbf{w}) = \text{softmax}(\text{fc}(\mathbf{w})\text{fc}(\mathbf{v})^\top)\text{fc}(\mathbf{v}) \tag{3.7}$$

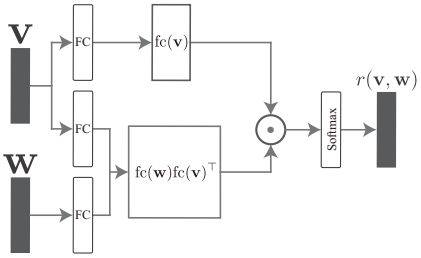


図 3.4: 類似性の示す値の算出過程

また，類似性に基づいた融合を式 (3.8) と式 (3.9) に示す．ただし，$\mathbf{z}_{\text{attn}}^{\text{early}}$ は early fusion 後のベクトル，$\mathbf{z}_{\text{attn}}^{\text{late}}$ は late fusion 後のベクトルとする．

$$\mathbf{z}_{\text{attn}}^{\text{early}} = [r(\mathbf{x}^{\text{a}^*}, \mathbf{x}^{\text{t}^*}), r(\mathbf{x}^{\text{t}^*}, \mathbf{x}^{\text{a}^*})] \tag{3.8}$$

$$\begin{aligned}\mathbf{z}_{\text{attn}}^{\text{late}} = [&r(\mathbf{y}^{\text{early}}, \mathbf{y}^{\text{a}}), r(\mathbf{y}^{\text{a}}, \mathbf{y}^{\text{early}}),\\ &r(\mathbf{y}^{\text{t}}, \mathbf{y}^{\text{early}}), r(\mathbf{y}^{\text{early}}, \mathbf{y}^{\text{t}}),\\ &r(\mathbf{y}^{\text{a}}, \mathbf{y}^{\text{t}}), r(\mathbf{y}^{\text{t}}, \mathbf{y}^{\text{a}})]\end{aligned} \tag{3.9}$$

Early fusion のみを用いた場合（No. 1–4），early fusion 結果を入力とする感情認識から得られた $\mathbf{y}^{\text{early}}$ をソフトマックス関数に入力した値が最終的な予測結果となる．Early fusion と late fusion を併用した場合（No. 5–20），3 種類の予測ベクトルの late fusion 結果を入力とする感情認識から得られた $\mathbf{y}^{\text{late}}$ をソフトマックス関数に入力した値が最終的な予測結果となる．Late fusion のみを使用した場合（No. 21–24）は，2 種類の予測ベクトルの late fusion 結果を入力とする感情認識から得られた $\mathbf{y}^{\text{late}}$ をソフトマックス関数に入力した値が最終的な予測結果となる．尚，和

による融合を late fusion に利用した場合は，融合したベクトルを直接ソフトマックス関数に入力する．

本研究では No. 9, 10 以外の early fusion と late fusion を併用した手法（No. 5–8，11–20）を提案手法とする．また，従来手法（No. 2–4, 9, 10, 21, 22）と合わせて early fusion のみ，または late fusion のみの実験（No. 1, 23, 24）を行ない，early fusion または late fusion と 4 種類の融合処理を組み合わせた全 24 通りの手法より得られた認識結果を比較する．

3.3 評価実験

本章では，第 3.2 章で説明した手法を用いた音声感情認識の学習，検証，評価を行い各手法の認識性能を比較する．また，評価時に正解テキスト，または音声認識結果を利用した場合の結果を比較する．

3.3.1 データセット

データセットには日本語感情音声データ JTES (Japanese twitter based emotional speech) [44] を使用した．四つの感情カテゴリ（喜び，悲しみ，怒り，平静）について，それぞれ 50 文，計 200 文を読み上げ文が用意されている．音声データは男性 50 名，女性 50 名の計 100 名分収録されている．

音響情報のみを利用した音声感情認識の評価では，データセットを話者・発話文の重複を許さないようにランダムに分割し，全ての話者の発話が学習と検証，評価に利用されるように 25 分割交差検証法で評価した．学習データには 20,000 発話の音声を利用した．

言語情報のみを利用したテキスト感情認識の評価ではデータセットを発話文の重複を許さないようにランダムに分割し，全ての発話文が学習と検証，評価に利用されるように 5 分割交差検証法で評価した．尚，言語情報のみを利用したテキスト感情認識の学習時にはデータ不足を補うためのデータ拡張を行なった．JTES の正解テキストを外国語に翻訳し，日本語に再翻訳する逆翻訳でデータを拡張した．翻訳対象の外国語には中国語，英語，スペイン語など言語人口が比較的多い言語を採用し，Google Translate Ajax API と DeepL API で逆翻訳した．逆翻訳結果の重複文は集約し，最終的に「怒り」のテキストが 606 文,「喜び」のテキストが 559 文,「喜び」のテキストが 572 文,「喜び」のテキスト 549 文，合計で 2,286 文になるようにデータを拡張した．尚，逆翻訳文と正解テキストが学習，検証，評価で重複しないように設定した．また，評価にはデータ拡張前の正解テキストのみを使用した．

音響・言語情報の融合に基づく音声感情認識の評価では，事前学習時と同様に話者・発話文の重複を許さない 25 分割交差検証法で評価した．学習データには 200 文の発話文と 20,000 発話の音声を利用した．

また，本研究では，音声認識から得られた書き起こしテキストを利用した評価も行なった．音

表 3.3: 音響情報のみを利用した音声感情認識のネットワーク構造

層の種類	パラメータ
Convolution1	channel (in):1, channel (out):32, kernel:(12, 5), stride:(2, 2)
Max pooling1	kernel:(2, 2), stride:(2, 2)
Convolution2	channel (in):32, channel (out):64, kernel:(8, 4), stride:(1, 1)
Max pooling2	kernel:(2, 2), stride:(2, 2)
Convolution3	channel (in):64, channel (out):128, kernel:(5, 2), stride:(1, 1)
Max pooling3	kernel:(2, 2), stride:(2, 2)
BLSTM	dim (in):128, dim (out):256, dim (hide):128, 2layers
Self-attention	dim (in):256, dim (out):256, 4heads
全結合層	dim (in):256, dim (out):64
全結合層	dim (in):64, dim (out):4

声認識には ESPNet [45] が提供する Transformer モデルを利用した．尚，使用する音声認識モデルは，日本語話し言葉コーパス [46] で事前に学習されている．音声認識から得られた書き起こしテキストを用いて，言語情報のみを利用したテキスト感情認識または音声・言語情報の融合に基づく音声感情認識の評価を行なった．尚，使用する事前学習済み音声認識モデルを JTES の音声で評価した場合の単語誤り率は 37.0% であった．

3.3.2 実験条件

本研究では音響情報と言語情報の融合のために，あらかじめ音響情報のみを利用した音声感情認識と言語情報のみを利用したテキスト感情認識を構築する．以降，各感情認識器の学習条件を説明する．

音響情報のみを利用した音声感情認識

音響情報のみを利用した音声感情認識には CNN と BLSTM，self-attention を組み合わせた構造を用いた．表 3.3 に詳細なネットワーク構造を示す．入力部では音響特徴量を 1 セグメントあたり 100 フレームになるよう分割し，1 セグメントごとに CNN へ入力を繰り返した．CNN から得られた出力はチャネル数を固定して時間次元と周波数次元を平坦化した．この処理で 1 セグメント当たり 128 次元のベクトルが 80 個得られる．これをセグメントの数だけ連結して得られた 80 個 × セグメント数の 128 次元のベクトルを順々に BLSTM へ入力した．BLSTM から得られた情報は self-attention に入力し，self-attention からの出力情報は行方向に和を算出した後に全結合層に入力した．最終的に入力発話に対して一つの予測結果が得らえる．尚，ネットワークの途中には活性化関数とドロップアウトを適宜挿入した．入力音声のサンプリングレートは 16,000Hz，データ形式は wav とした．入力特徴量は 400 次元の対数パワースペクトログラム（10 Hz grid resolution）[47]

とする．このときの離散フーリエ変換の点数は 1,600，フレーム幅は 40ms，シフト幅は 10 ms とした．離散フーリエ変換より得られた特徴量の内，0Hz から 4,000Hz までの特徴のみを入力特徴量として利用した．抽出した特徴量には各特徴量次元の値が平均 0，分散 1 になるように正規化を行った．エポック数は 100，バッチサイズは 16 とした．最適化手法には Adam [48] を採用し，warm up を適用した．また，学習率は 0.0001，損失関数は cross entropy とした．尚，評価時には，全エポックの中で検証誤差が最も小さかったモデルパラメータを用いた．

言語情報のみを利用したテキスト感情認識

言語情報のみを利用したテキスト感情認識には BERT と全結合層を用いた．本研究では Transformers [49] が提供している BERT のモデル構造を採用した．BERT の事前学習済みモデルには日本語版 Wikipedia の大規模テキストで学習したパラメータを利用した．BERT の事前学習済みパラメータは予め固定し，最終層の線形層のみを学習させた．入力テキストの形式は，単語 ID ベクトルとした．ID 化には MeCab [50] に基づく Transformers の tokenizer を使用した．エポック数は 300，バッチサイズは 128 とした．最適化手法には Adam を採用し，warm up を適用した．また，学習率は 0.0001，損失関数は categorical cross entropy とした．損失の計算時には各感情カテゴリにおけるデータ数の不均衡を考慮するため，学習時に用いたデータ数の逆数を掛け合わせた．よって，実際に計算する損失関数 L は式 (3.10) になる．ただし，ネットワークの最終出力から得られる事後確率を p，正解ラベルの値を q，学習データ数を U とする．また，感情カテゴリ数を K，各感情カテゴリの添え字を k とする．

$$L = -\sum_{k}^{K} \frac{1}{U_k} q_k \log\left(p_k\right) \tag{3.10}$$

尚，評価時には，全エポックの中で検証誤差が最も小さかったモデルパラメータを用いる．

音響・言語情報の融合に基づく音声感情認識

音響・言語情報の融合に基づく音声感情認識には 3.3.2 節で利用した CNN と BLSTM，self-attention，3.3.2 節で利用した BERT と全結合層を用いた．JTES で事前学習済みの音響情報のみを利用した音声感情認識と言語情報のみを利用したテキスト感情認識を組み合わせた．各認識モデルのパラメータは予め固定した．

音響情報を利用した感情認識モデルから得られる中間ベクトルの次元数は 256 次元，言語情報を利用した感情認識モデルから得られる中間ベクトルの次元数は 768 次元とした．Early fusion 及び late fusion で得られた値を入力とする感情認識は全結合層から構成されている．Early fusion 時には各中間ベクトルを一度全結合層に入力し得られた値を融合した．尚，ネットワークの途中には活性化関数とドロップアウトを適宜挿入した．各認識モデルと early fusion から得られる予測ベクトルの次元数は 4 次元とした．エポック数は 100，バッチサイズは 16 とした．最適化手法には

表 3.4: 音響情報のみ/言語情報のみを利用した音声感情認識の正解率

	各感情の正解率				Ave.
	ang.	joy	sad.	neu.	
音響情報のみ	66.5	25.1	51.2	**72.4**	53.8
言語情報のみ（(a) 評価：正解テキスト）	**72.0**	**70.0**	**54.0**	30.0	**56.5**
言語情報のみ（(b) 評価：音声認識結果）	70.8	49.0	36.2	35.2	47.8

Adam を採用し，warm up を適用した．学習率は 0.0001，損失関数は cross entropy とした．尚，評価時には，全エポックの中で検証誤差が最も小さかったモデルパラメータを用いる．

3.3.3 実験結果

音響情報のみ/言語情報のみを利用した場合

音響情報のみを利用した音声感情認識と言語情報のみを利用したテキスト感情認識の実験結果を表 3.4 に示す．ただし，(a) 及び (b) は評価データとしてそれぞれ正解テキスト，音声認識結果を用いた場合の感情の正解率を示す．音響情報のみの正解率の平均は言語情報のみの正解率の平均と比較して約 2.7% ポイント認識率が低い．特に，音響情報のみを利用した場合，喜びの正解率が非常に低い．これは，言語情報のみを利用したテキスト感情認識の方が，文章中に感情を表す単語が表出した場合，比較的感情を決めやすいためだと考えられる．一方で，平静の発話には特徴的な単語が含まれないため，音声の音響的な変化から決める音響情報を利用する場合の方が正解率が高くなったと考えられる．

言語情報のみを利用したテキスト感情認識について，評価に正解テキストを利用した場合と音声認識結果を利用した場合の結果を比較すると，音声認識結果を利用する場合の方が全体的に正解率が低い．原因として音声認識の単語誤り率が高いことが挙げられる．既存の学習済み音声認識モデルでは感情音声を上手く書き起こせず，感情を決める単語等を感情認識に入力出来ていないと考えられる．一方で，感情が上手く決定できない場合が増加し全体的に平静と認識するようになったため，平静の正解率が高くなったと考えられる．

音響情報と言語情報を融合した場合

音響・言語情報の融合に基づく手法の実験結果を表 3.5 に示す．ただし，No. は表 3.1 で示した各手法の ID 番号である．表 3.4 と表 3.5 の結果を比較すると，殆どの手法で平均正解率が向上している．特に音声認識結果で評価した場合，全ての手法において言語情報のみの手法よりも平均認識率が向上している．従って，従来研究と同様に音響情報と言語情報を併用することで音声感情認識の正解率が向上することが示せた．

表 3.5: 音響・言語情報の融合に基づく音声感情認識の正解率

No.	(a) 評価：正解テキスト					(b) 評価：音声認識結果				
	各感情の正解率				Ave.	各感情の正解率				Ave.
	ang.	joy	sad.	neu.		ang.	joy	sad.	neu.	
1	70.8	59.5	70.3	77.9	69.6	69.4	52.8	62.3	78.9	65.9
2 [20]	72.2	59.1	66.8	**79.4**	69.4	70.2	51.5	59.7	**80.2**	65.4
3 [42]	69.1	51.8	65.3	78.2	66.1	69.4	43.1	59.3	78.8	62.6
4 [23]	70.6	39.5	63.9	65.8	59.9	69.4	37.0	60.4	65.1	58.0
5 [Ours]	66.1	57.6	70.4	77.1	67.8	65.2	51.0	62.2	78.4	64.2
6 [Ours]	70.2	59.7	72.4	78.9	**70.3**	67.6	54.1	64.2	79.6	**66.4**
7 [Ours]	62.4	53.3	65.2	71.8	63.2	61.0	48.2	57.3	74.7	60.3
8 [Ours]	67.9	56.8	72.3	71.1	67.0	67.7	49.6	62.5	75.2	63.7
9 [40]	74.6	55.3	68.0	74.6	68.1	74.8	48.7	59.3	77.3	65.0
10 [41]	70.1	60.9	**74.8**	70.8	69.1	67.0	55.8	**67.0**	71.9	65.4
11 [Ours]	64.6	55.9	66.8	70.2	64.4	63.4	52.6	58.0	72.2	61.6
12 [Ours]	65.4	54.3	74.1	70.2	66.0	65.7	45.5	63.3	73.8	62.1
13 [Ours]	74.9	52.9	71.8	71.5	67.8	74.1	44.4	61.6	74.5	63.6
14 [Ours]	66.5	39.4	64.4	77.6	62.0	65.7	34.5	57.5	77.6	58.9
15 [Ours]	55.9	37.1	64.9	60.8	54.7	55.5	32.1	56.8	62.0	51.6
16 [Ours]	70.1	44.0	47.4	42.9	51.1	69.3	38.5	44.0	45.1	49.2
17 [Ours]	74.1	54.5	70.3	68.3	66.8	74.0	44.1	60.5	73.1	62.9
18 [Ours]	65.7	50.2	49.4	59.2	56.1	63.7	49.3	47.1	60.8	55.2
19 [Ours]	65.4	54.9	62.8	57.5	60.2	66.4	50.1	58.1	61.2	58.9
20 [Ours]	55.8	34.9	72.2	52.7	53.9	54.1	30.6	63.7	57.8	51.6
21 [43]	**82.1**	59.1	71.5	64.3	69.2	**83.3**	48.2	61.7	69.2	65.6
22 [39]	68.7	**70.0**	71.3	45.1	63.8	67.9	50.4	57.6	51.4	56.8
23	52.7	52.0	40.5	57.3	50.6	53.8	49.4	35.7	60.3	49.8
24	65.0	63.2	67.5	54.1	62.4	65.8	**57.5**	59.9	56.2	59.9

4 種類の融合処理における平均正解率を比較すると，early fusion のみ（No. 1），または late fusion のみ（No. 21）の両方の手法で和による融合を利用した場合が最も正解率が高い．一方で，直積による融合や類似性に基づいた融合を利用した場合（No. 3, 4, 23, 24）は全体的に正解率が低くなった．理由として early fusion で結合する中間ベクトルの各要素の関連性を考慮することに効果が無かったことが考えられる．Late fusion も同様の理由で正解率の改善には大きな効果が無かったと言える．Early fusion と late fusion を併用した手法を比較すると，early fusion には和

による融合，late fusion には結合による融合を選択した手法（No. 6）が最も平均正解率が高かった．従来手法（No. 10）と比較すると，約 1.2% ポイントの正解率の向上が見られた．尚，符号検定 [51] によって従来手法に対する提案手法（No. 6）の有効性を調べた結果，提案手法で組み合わせた ealy fusion のみの手法（No. 1）と late fusion のみの手法（No. 22）に対してそれぞれ有意水準 10%と 5%で有意差が見られた．また，従来手法 (No. 10) に対しても有意水準 5%で有意差が見られた．

評価に正解テキストを利用した場合と音声認識結果を利用した場合の結果を比較すると，入力が音声認識結果の場合の方が正解率が低いことが分かる．また，正解テキストを利用した場合よりも音声認識結果を利用した場合の方が怒りと喜び，悲しみの正解率が低く，平静の正解率が全体的に高いことが分かる．故に，言語情報のみを利用したテキスト感情認識と同様に，音声認識による書き起こしの誤りによって感情を上手く決定できなかったため，全体的に発話を平静に分類するようになったと考えられる．

以上より，音響情報のみ，または言語情報のみを利用した音声感情認識よりも，音響・言語情報の融合に基づく音声感情認識の方が正解率が高いことが分かった．特に，提案手法（No. 6）が最も正解率が高いことが分かった．

3.4 まとめ

本研究では early fusion と late fusion を併用した手法を提案し，4 種類の融合処理を用いて様々な手法を比較した．結果は，early fusion に和による融合，late fusion に結合による融合を利用する組み合わせが，正解率の向上に最も効果的であることが明らかになった．本研究では各認識器の中間ベクトルと予測ベクトルがいずれも固定長ベクトルであることから，明示的に関連性を最適化するような直積や類似性に基づいた融合はあまり効果的ではなかった可能性がある．また，音声認識結果を用いた評価では，平静の正解率が向上し，その他の感情の正解率が低下した．言語情報を利用する場合，音声認識結果の誤りが音声感情認識の感情ごとの正解率に影響することが分かった．

第4章　感情ラベル列を用いた音声感情認識

4.1　はじめに

感情ラベル列を用いた音声感情認識では認識器の学習時にどのように感情ラベル列を与えるかが課題である．従来研究では，発話内の有声音素には感情が表出し，無声音素などには感情が表出しないという仮定に基づいた感情ラベル列を利用し認識器を学習していた．そのため，入力音声の母音や有声子音で音響的な差異があったとしても，母音の発話区間と有声子音の発話区間は同じ感情クラスとして分類され，無声子音は例外なく非感情クラスとして分類されていた．しかし，本来音響的に大きな差異があるならば別クラスとして分類する方が適切である．また，関連研究 [52–54] では母音や有声子音，無声子音，記号によって感情表出が異なることが示されているため，大きな差異がある音素クラスで感情状態を定義した方が，より細かな感情認識が可能になると考えた．

本研究では、音声の細かな音響的差異を考慮した新しいフレーム単位の音声感情認識手法を提案する．音素に依存する音響的な多様性を考慮するため，各音素の属性記号（音素クラス属性）を付与した感情ラベル列を用いて認識器を学習する．また，既存研究で評価されていなかったフレーム単位の感情認識の精度を，発話内で感情が変化する音声を用いて評価した．結果より，フレーム単位の音声感情認識モデルはフレーム単位で感情を認識することができ，提案手法によってその性能が改善することが分かった．

4.2　従来の感情ラベル列を用いた音声感情認識

感情ラベル列を用いた音声感情認識では，感情ラベル列を数フレーム単位で逐次的に推定する．従来研究 [26] では，有声音素は感情状態，無音区間や無声音素などには無感情状態 (NULL) を仮定し，発話単位の感情ラベルから感情ラベル列を構築している．感情ラベル列を用いた音声感情認識の学習と予測の概要を図 4.1 に，従来手法における書き起こし文を感情ラベル列に変換する手順を図 4.2 に示す． 尚，従来手法の概要は赤背景の箇所に示している．学習時では，はじめに入力発話に対応する書き起こしテキストを音素に変換し，母音と有声音の合計数だけ発話単位の感情ラベルを並べることで感情ラベル列を構築している．例えば，喜び（happiness; H）の “YES, YES. [LAUGHTER]” (/jɛs, jɛs. [LAUGHTER]/) という発話の場合，有声子音（/j/）が 2 つと母音（/ɛ/）が 2 つ含まれているため，感情ラベル列は {H, H, H, H} のようになる．この感情ラベル列を用いた手法には様々なニューラルネットワークモデルが利用されている．例えば，BLSTM

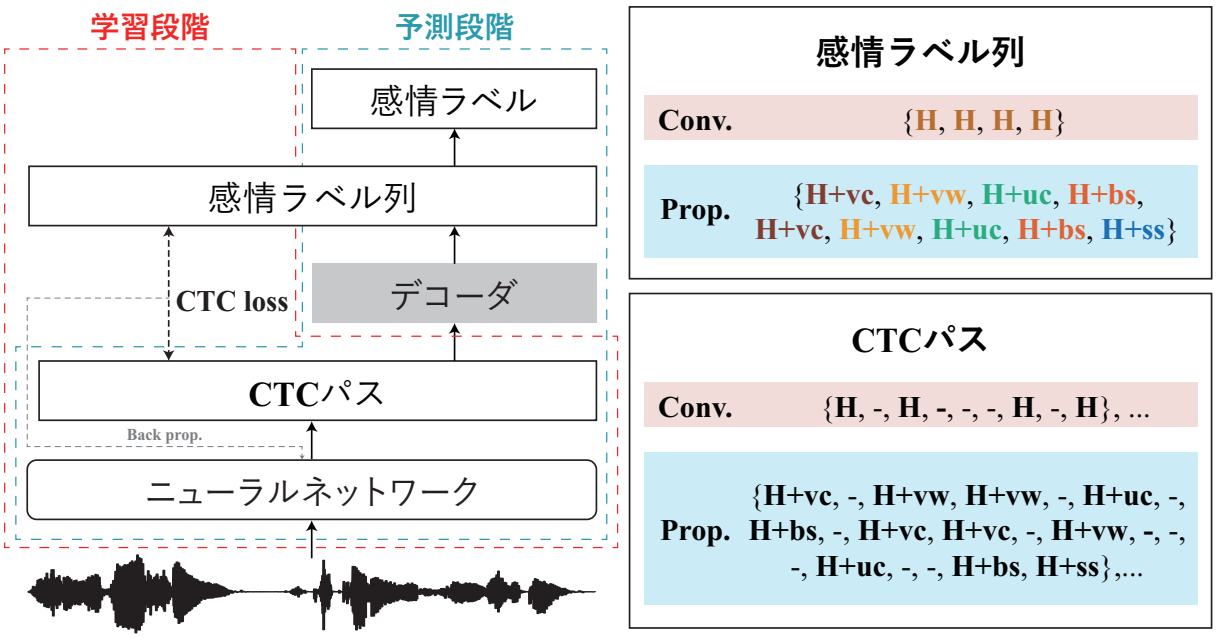


図 4.1: 感情ラベル列を用いた音声感情認識の学習及び予測の概要

図 4.2: ラベルの作成方法

と様々な注意機構を組み合わせたモデル [55] や parallel CNNs と SENet (squeeze-and-excitation network)，DRN (dilated residential network) を組み合わせたモデル [56] などがある．

推定モデルは CTC (connectionist temporal classification) [57] に基づいて学習される．CTC とは，どのクラスにも該当しない blank 記号 (-) と記号の繰り返しが含まれるパス (CTC パス) の推定する枠組みである．尚，blank 記号は無感情状態に該当する．この手法では，入力系列長より出力系列長が短い場合であっても出力系列を推定できる．CTC に基づく学習では，長さ T の入力系列 $\mathbf{x} = [x_0, \ldots, x_{\mathrm{T}}]$ が与えられたとき，長さ L($\leq$ T) の感情ラベル列 $\mathbf{y} = [y_0, \ldots, y_{\mathrm{L}}]$ が得られる確率 $p(\mathbf{y}|\mathbf{x})$ を最大化する．求める確率 $p(y|x)$ と CTC の損失関数をそれぞれ式 (4.1) と式 (4.2) に

示す．ただし，x_t は時刻 t の入力フレーム，π_t は時刻 t の感情ラベル，π は感情ラベル列の CTC パス，$\Phi(y)$ は π の集合を示す．また，U は学習データの集合を示す．

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\pi \in \Phi(\mathbf{y})} \prod_{t=1}^{\mathrm{T}} p(\pi_t|x_t) \tag{4.1}$$

$$\mathcal{L}_{\mathrm{ctc}} = -\sum_{(\mathbf{x},\mathbf{y}) \in U} \log p(\mathbf{y}|\mathbf{x}) \tag{4.2}$$

各時刻で予測するクラス数は，予測する感情状態数 +1（blank 記号）である．評価時には，出力の CTC パスに対して blank 記号の削除と連続する同一感情ラベルの集約を適用することで予測感情ラベル列が得られる．最終的に，CTC パスをフレーム単位の予測結果とし，予測感情ラベル列の中で最も登場頻度が多い感情カテゴリを発話単位の予測結果とする．

4.3 音素クラス属性付き感情ラベル列を用いた音声感情認識

音素クラス属性を考慮した感情ラベル列を推定する手法を提案する．本研究では，基本記号（basic symbols; bs），母音（vowels; vw），有声子音（voiced consonants; vc），無声子音（unvoiced consonants; uc），特殊記号（special symbols; ss）の 5 つの音素クラスを定義した．基本記号は句読点や感嘆符などを示し，特殊記号はデータセット毎に用意された独自の記号などを示す．Lee らは母音が音声感情認識において重要な情報であると示している [52]．また，Aryani らは有声・無声子音も様々な感情を表現している可能性を示唆している [53]．基本記号や特殊記号については，これらの埋め込み表現を入力特徴量に使用することで認識性能が向上することが従来研究 [23, 54] で示されている．そのため，提案手法では母音や有声・無声子音，その他の記号の属性を考慮した．以上の属性に該当しない記号は非感情状態とする．図 4.2 の青背景の箇所に示す通り，提案手法では音素クラス属性と感情ラベルを組み合わせて感情ラベル列を構築し，図 4.1 の青背景の箇所に示す通り，CTC モデルの学習に利用する．従来手法との大きな違いは，様々な種類の音素属性の情報を明示的に考慮して感情ラベル列の推定モデルを学習できる点である．例えば，喜び（Happiness; H）の “YES, YES. [LAUGHTER]”（/jɛs, jɛs. [LAUGHTER]/）という発話の場合，基本記号（/,/, /./）を 2 つ，母音（/ɛ/）を 2 つ，有声音素（/j/）を 2 つ，無声音素（/s/）を 2 つ，特殊記号（/[LAUGHTER]/）を 1 つ含むため，感情ラベル列は {H+vc, H+vw, H+uc, H+bs, H+vc, H+vw, H+uc, H+bs, H+ss} のようになる．“感情ラベル + 音素クラス属性” は音素クラスを持つ感情ラベルを示す．CTC パスの各フレームで推定されるクラス数は，感情クラス数 × 音素クラス属性数 +1（blank 記号）である．学習時には，音素クラス属性をもつ感情ラベル列を利用して deep neural network(DNN) モデルを学習する．評価時には，推定された CTC パスをフレーム単位の予測結果とし，CTC パスから得られる予測感情ラベル列の内，最も出現頻度の高い感情を発話単位の予測結果としてモデルの性能評価を行う．

表 4.1: 感情別の音声データ数

Session	Ang.	Hap.	Sad.	Neu.	Total
1	62	132	104	223	521
2	22	191	100	217	530
3	90	149	190	198	627
4	84	195	81	174	534
5	31	280	133	287	731
Total	289	947	608	1099	2943

4.4 実験条件

本節では実験条件について述べる．まず，感情ラベル列を用いた従来方法と提案方法で学習したモデルについて，発話単位の認識率を比較した．また，結果を先行研究 [26, 55, 56] で報告された結果とも比較した．更に，提案方法がフレーム単位の性能を改善できるかどうかを調べるために，異なる感情が含まれる発話を用いてフレーム単位の認識率を比較した．実験で使用したデータセットとモデル，評価指標の詳細については次に述べる．

4.4.1 データセット

データセットには英語の感情音声が収録された IEMOCAP (Interactive emotional dyadic motion capture database) [58] を用いた．IEMOCAP は対話中の感情音声を収録したデータセットである．5 つのセッションで構成されており，各セッション男女 2 名の対話が収録されている．また，台本を読み上げた演技音声が 5,255 発話，台詞無しの脚本に沿った即興対話音声が 4,784 発話，計 10,039 発話が収録されている．感情の正解ラベルには,「喜び」「悲しみ」「怒り」「不満」「興奮」「失望」「平静」「驚き」「その他」の 10 種類が付与されている．

発話単位の認識性能の評価では，評価実験では従来研究に従い即興対話音声のみを利用し,「喜び」「悲しみ」「怒り」「平静」の 4 感情分類を行なった．尚,「喜び」の音声データは「興奮」の音声データを追加して使用した．学習・検証・評価に使用したデータの詳細は表 4.1 に，発話長別の音声データ数は図 4.3 に示す．学習に使用した音声データの平均時間長は約 4.5 秒である．学習時には，従来研究 [14, 59] を参考に 15 秒以下の音声を使用した．

フレーム単位の認識性能の評価では，発話単位の評価と同じ学習データで作成した発話内で感情ラベルが変化する評価データを利用した．評価データの作成手順を図 4.4 に示す．初めに 2 つの異なる感情カテゴリからそれぞれ 1 つの発話を選択した．いずれの発話の長さは平均発話長よりも長いものとする．本研究では認識対象の感情は 4 カテゴリあるため，4 カテゴリの順列は 12 ペアあった．次に ctc-segmentation [60] によって各発話の発話期間とそれに対応する単語の対応を取り，発話毎の総単語数の半分になるように各発話を分割した．この処理により発話が単語の途

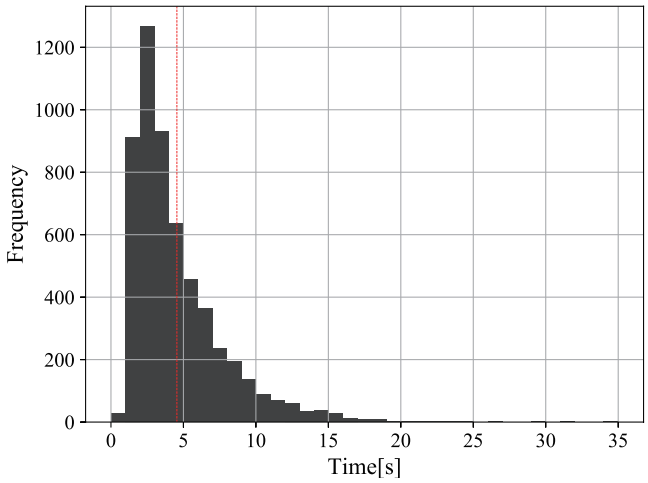


図 4.3: 発話長別の音声データ数（赤線：平均時間長）

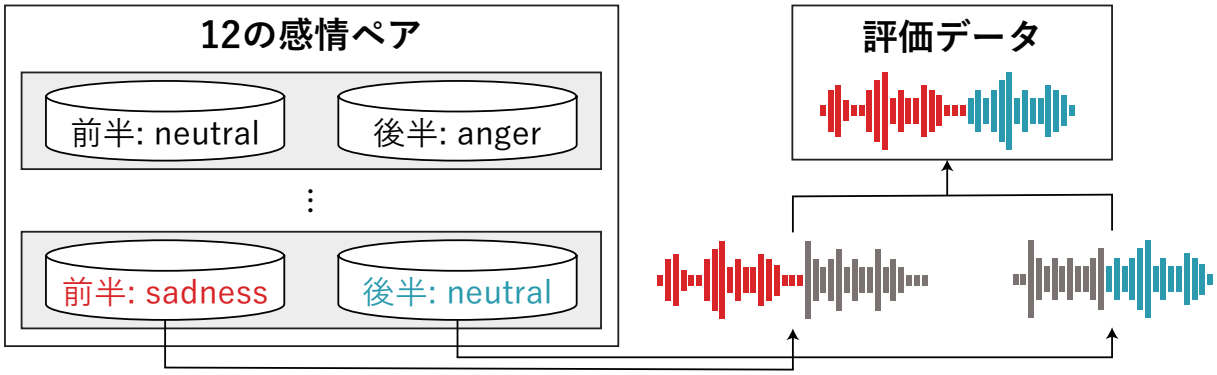


図 4.4: 評価データの作成方法

中で切れないようにした．最後に各ペアの最初の感情カテゴリの部分発話と 2 番目の感情カテゴリの部分発話を結合した．尚，各 fold で 600 個の組み合わせ発話を用いてフレーム単位の認識性能を評価した．

感情ラベル列の構築に必要な音素列は書記素音素変換ツールキット[1]によって取得した．音素クラス属性と記号の対応関係は表 4.2 に示す．IEMOCAP に収録されているテキストにはそれぞれ基本記号が 37,304 個，母音が 40,474 個，有声子音が 37,388 個，無声子音が 21,791 個，特殊記号が 133 個含まれている．本実験では，提案手法の有効性を調べるために異なる音素クラス属性を考慮した 3 種類の手法を評価した．表 4.3 は，従来手法及び提案手法で考慮されていた音素クラス属性と CTC パスの各フレームで推論されるクラス数を示している．考慮する音素クラス属性以外は非感情状態を示す記号（blank 記号）とする．評価は話者重複なしの 10 分割交差検証法で行なった．尚，異なる初期値でモデルは 5 回 10 分割交差検証を実施しており，それらの評価結果の平均で認識性能を評価している．各 fold では 8 名の音声を学習データ，1 名の音声を検証データ，1 名の音声を評価データとして利用した．発話単位の評価では，各 fold の評価データを用いた．一方，フレーム単位の評価では，第 4.4.1 で作成した感情の変化を伴う評価データを各 fold で用いた．

[1]https://github.com/Kyubyong/g2p

表 4.2: 音素クラス属性と記号の対応関係

音素クラス属性	記号
基本記号 (bs)	!, ?, ’, ,, -, ., >
母音 (vw)	AA, AE, AH, AO, AW, AY, EH, ER, RY, IH, IY, UH, UW, OW, OY
有声子音 (vc)	B, D, DH, G, L, M, N, NG, JH, R, V, W, Y, Z, ZH
無声子音 (uc)	CH, F, HH, K, P, S, SH, T, TH
特殊記号 (ss)	[LAUGHTER], [LIPSMACK], [GARBAGE], [BREATHING]

表 4.3: 各手法で推論されるクラス数

	考慮する音素クラス属性	推定クラスの数
Conv.	voiced	5 (4 感情 + 1 blk.)
Prop. I	vw, vc, ss	13 (4 感情 × 3 属性 + 1 blk.)
Prop. II	vw, vc, uc, ss	17 (4 感情 × 4 属性 + 1 blk.)
Prop. III	bs, vw, vc, uc, ss	21 (4 感情 × 5 属性 + 1 blk.)

4.4.2 推定モデルの学習条件

感情ラベル列の推定モデルには，事前学習済みの wav2vec2.0 と HuBERT を利用した．wav2vec2.0 は自己教師あり表現学習手法の一つである [61]．この手法では，マスクされていないフレームを負例とし，マスクされたフレームの音声表現を推定するように対照学習が行われる．HuBERT も同様に自己教師あり表現学習手法の一つである．この手法では，フレーム単位の音声表現をいくつかのクラスに分類することでモデルを学習する．尚，分類クラスは，音響特徴量を用いたフレーム単位のクラスタリング結果に基づいて定義される．本実験では，Hugging Face [49] が提供する wav2vec2.0 及び HuBERT の事前学習済みモデル (`facebook/wav2vec2-large-960h-lv60`, `facebook/hubert-large-ls960-ft`) を利用した．これは，英語音声データセットの Libri-Light [62] と Librispeech [63] を用いて音声表現の事前学習と自動音声認識の fine-tuning を行なったモデルである．実験に使用したモデル構造は，7 層の CNN と 24 層の Transformer block からなる wav2vec2.0 または HuBERT と 1 層の全結合層で構成されている．学習時には CNN 層のモデルパラメータを固定し，後段の Transformer block と全結合層を学習した．モデルの入力は音声波形であり，出力は感情ラベル列または音素属性を考慮した感情ラベル列の CTC パスである．エポック数は 50, バッチサイズは 8，学習率は 0.0001，最適化手法には RAdam [64] を利用した．また，閾値が 5.0 の gradient clipping を適用した．損失関数は CTC loss である．

4.4.3 推定モデルの評価指標

発話単位の評価指標には，WA (weighted accuracy) と UA (unweighted accuracy) を利用した．WA と UA はそれぞれ式 (4.3) と式 (4.4) で算出される．ただし，評価データ数を N_{all}，分類クラス数を K，k 番目の分類クラスの評価データ数を $N_{\mathrm{all},k}$，k 番目の分類クラスの正解数を N_k とする．また N_k は式 (4.5) で求められる．ただし，n 番目の評価データにおける正解感情 ID を z_n，予測感情 ID を $\hat{z}_n$，$\mathbb{1}$ を指示関数とする．WA は全体の正解率，UA は各感情の再現率の平均を示している．尚，何も予測され無かった発話の感情は平静として評価する．WA および UA が高くなるほど，モデルは発話単位の感情を正確に予測できることを示す．

$$\mathrm{WA} = \frac{1}{N_{\mathrm{all}}} \sum_{k=1}^{K} N_k \tag{4.3}$$

$$\mathrm{UA} = \frac{1}{K} \sum_{k=1}^{K} \frac{N_k}{N_{\mathrm{all},k}} \tag{4.4}$$

$$N_k = \sum_{n=1}^{N_{\mathrm{all},k}} \mathbb{1}(z_n = \hat{z}_n) \tag{4.5}$$

フレーム単位の評価指標には，EMR (emotion match rate) を利用した．EMR は式 (4.6) で算出される．ただし，評価データの全フレーム数を M_{all}，感情状態の集合を $\mathbb{U}_{\mathrm{emo}}$，$m$ 番目のフレームにおける正解感情 ID を a_m，予測感情 ID を $\hat{a}_m$ とする．EMR は感情状態にあると予測された全フレームの内，正しく予測されたフレームの割合を示している．尚，予測トークンが black 記号の場合，そのフレームは非感情状態とする．EMR が高くなるほど，モデルはフレーム単位の感情を正確に予測できることを示す．

$$\mathrm{EMR} = \frac{\displaystyle\sum_{m=1}^{M_{\mathrm{all}}} \mathbb{1}((\hat{a}_m \in \mathbb{U}_{\mathrm{emo}}) \wedge (a_m = \hat{a}_m))}{\displaystyle\sum_{m=1}^{M_{\mathrm{all}}} \mathbb{1}(\hat{a}_m \in \mathbb{U}_{\mathrm{emo}})} \tag{4.6}$$

また，従来手法と提案手法の有意差を調べるための検定も実施した．WA と UA を用いた発話単位の評価では,「手法間で誤認識数の差を算出し，正になる頻度と負になる頻度が等しくなった場合，手法間で結果に有意差がない」という帰無仮説を設定した．EMR を用いたフレーム単位の評価では,「手法間で EMR の差を算出し，正になる頻度と負になる頻度が等しくなった場合，手法間で結果に有意差がない」という帰無仮説を設定した．これらの仮説に基づいて、両側符号検定を行った．

表 4.4: 各手法における発話単位の評価結果

モデル	wav2vec2.0+FC		HuBERT+FC	
	WA (%)	UA (%)	WA (%)	UA (%)
Conv.	73.3	72.7	71.3	69.4
Prop. I	75.2*	74.3*	73.8*	72.2*
Prop. II	**75.5***	**74.5***	73.9*	72.3*
Prop. III	74.7*	73.9*	**74.4***	**72.7***

* 従来研究と比較したときの p 値が $p < 0.05$ になった場合

4.5 実験結果

4.5.1 発話単位の評価結果

表 4.4 は各手法における発話単位の WA 及び UA を示している．ただし，Conv. と Prop.I〜III は表 4.3 に対応している．各提案手法と従来手法の WA 及び UA との間には，$p < 0.05$ で有意差があった．

Conv. と Prop.I〜III を比較すると，全ての Prop. で Conv. よりも WA と UA が向上していることが分かる．Prop.I と Conv. を比較すると，Prop.I によって wav2vec2.0+FC では WA が 1.9%ポイント，UA が 1.6%ポイント向上，HuBERT+FC では WA が 2.5%ポイント，UA が 2.8%ポイント向上した．同じ有声音素であっても母音と有声子音を区別することで，推定モデルは音素毎の音響特性の差異に起因する細かな感情を認識できるようになったと考えられる．Prop. II と Conv. を比較すると，Prop. II によって wav2vec2.0+FC では WA が 2.2%ポイント，UA が 1.8%ポイント向上，HuBERT+FC では WA が 2.6%ポイント，UA が 2.9%ポイント向上した．考慮する音素クラス属性に無声子音を加えることで，推定モデルは発話中の呼吸速度やピッチの変化を示す無声音素に関連した感情を認識した可能性がある．Prop. III を Conv. を比較すると，Prop. III によって wav2vec2.0+FC では WA が 1.4%ポイント，UA が 1.2%ポイント向上，HuBERT+FC では WA が 3.1%ポイント，UA が 3.3%ポイント向上した．考慮する音素クラス属性に基本記号を加えることで，推定モデルは発話中の呼吸速度の細かな変化に関連した感情を認識した可能性がある．

Prop. I〜III を比較すると，wav2vec2.0+FC を用いた場合では Prop. II，HuBERT+FC を用いた場合では Prop. III で最も高い認識率を示した．推論モデルによって有効な提案手法が異なる原因としては，自己教師あり学習モデルの事前学習の方法の違いによるものと考えられる．wav2vec2.0 ではモデルを対照学習するのに対し，HuBERT のではフレーム毎のクラスタリング結果を用いて学習されている．そのため，フレーム分類で事前学習された HuBERT を利用した場合，最も細かく音素クラス属性を考慮する Prop. III が提案手法の中で最も高い WA と UA を達成したと考えられる．一方，マスクされたフレームとマスクされていないフレーム間の対照学習で事前学習され

表 4.5: 従来手法及び提案手法を用いたときの発話単位の認識率の比較

従来手法	WA (%)	UA (%)
BLSTM [26]	64.2	65.7
BLSTM+Component Attention [55]	69.0	67.0
PCNSE+SADRN [56]	73.1	66.3
提案手法		
HuBERT+FC (prop. III) [Ours]	74.4	72.7
wav2vec2.0+FC (prop. II) [Ours]	**75.5**	**74.5**

表 4.6: 各手法におけるフレーム単位の評価結果

モデル	EMR(%)	
	wav2vec2.0+FC	HuBERT+FC
Conv.	46.6	44.1
Prop. I	47.6*	45.1*
Prop. II	48.8*	46.9*
Prop. III	**49.0***	**47.0***

* 従来研究と比較したときの p 値が $p < 0.05$ になった場合

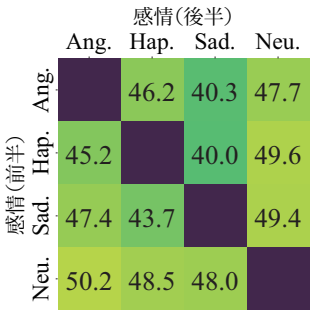
た wav2vec2.0 を利用した場合，基本シンボルを考慮する効果は明確ではなかった可能性がある．以上の結果から，自己教師あり学習時の正解信号や学習方法によって，考慮すべき音素クラス属性が変わる可能性があることが分かった．

学習に使用した自己教師あり学習モデル間の結果を比較すると，wav2vec2.0+FC を用いた場合は HuBERT+FC を用いた場合よりも WA と UA が高くなっていた．wav2vec2.0 は対照学習によって似たような音響特徴を持つフレームとそうでないフレームの差異を学習しているため，フレーム単位の音響的差異をより適切に識別できると考えられる．

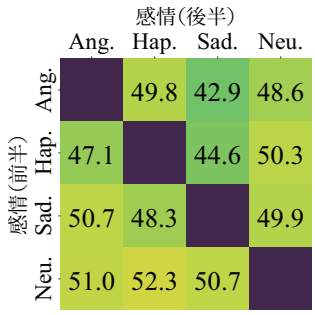
表 4.5 は，提案手法とこれまでの研究で報告されている WA と UA を示している．提案手法は，これまでの研究よりも同等かそれ以上の WA と UA を達成している．特に Prop. II で wav2vec2.0+FC を用いた場合，従来手法よりも最大で WA が 2.4%ポイント，UA が 8.2%ポイント改善した．この結果は，感情ラベル列を用いた音声感情認識においても自己教師あり学習モデルを fine-tuning することで性能を効果的に改善できることを示している．

4.5.2 フレーム単位の評価結果

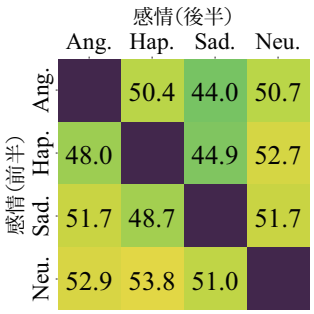
表 4.6 は，各手法におけるフレーム単位の EMR を示している．各提案手法と従来手法の EMR との間には，$p < 0.05$ で有意差があった．



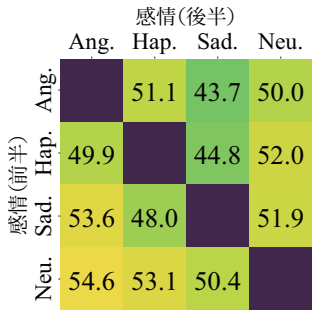
(a) Conv.



(b) Prop. I

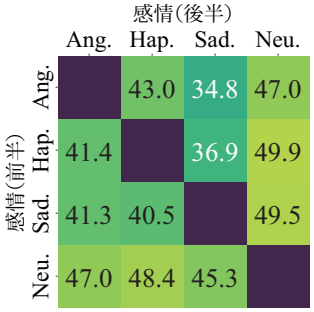


(c) Prop. II

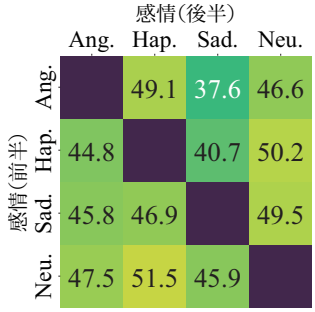


(d) Prop. III

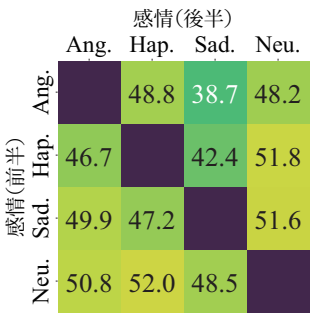
図 4.5: 評価音声の前半及び後半の EMR（wav2vec2.0+FC）



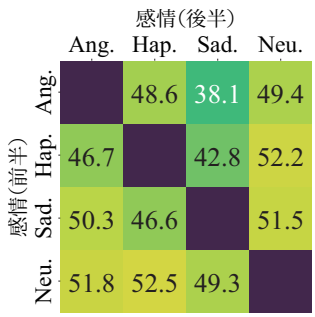
(a) Conv.



(b) Prop. I



(c) Prop. II



(d) Prop. III

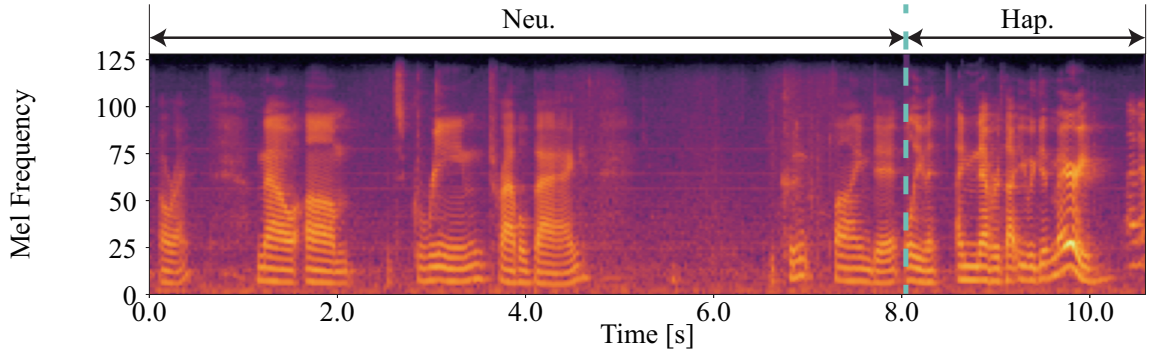
図 4.6: 評価音声の前半及び後半の EMR（HuBERT+FC）

Conv. と Prop. I〜III を比較すると，全ての Prop. で Conv. よりも EMR が向上していることが分かる．これは，感情ラベル列に音素クラス属性を組み込むことで，フレーム単位の音声感情認識のフレーム単位の認識性能を効果的に改善できることを示している．Prop. I〜III を比較すると，Prop. III が最も EMR が高いことが分かる．Prop. III の EMR は Conv. よりも wav2vec2.0+FC の場合で 2.4%ポイント，HuBERT+FC の場合で 2.9%向上している．これらは，音素クラス属性を母音や有声・無声子音，基本記号のように詳細に考慮することでフレーム単位の認識性能を改善できることを示している．学習に用いた各自己教師あり学習モデルの結果を比較すると，wav2vec2.0 を用いた場合の方が HuBERT を用いた時よりも EMR が高かった．これは第 4.5.1 節で述べた通り，wav2vec2.0 はフレーム同士の違いを区別するように学習されるため，フレーム単位の音声感情認識に適していると考えられる．

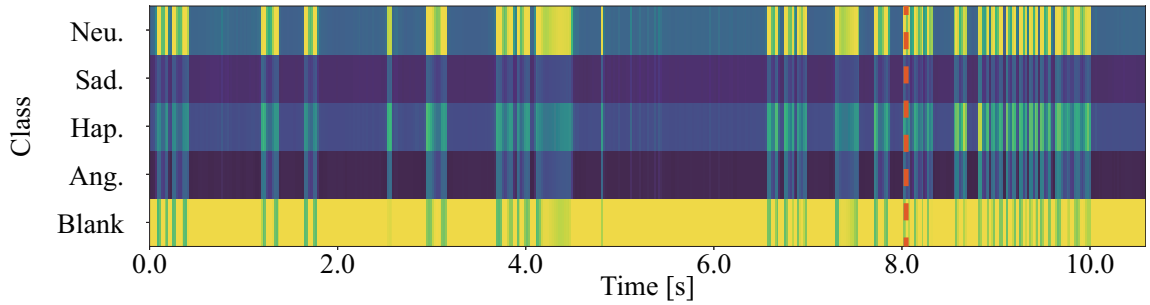
図 4.5 及び 4.6 は，wav2vec2.0+FC 及び HuBERT+FC を用いたフレーム単位の音声感情認識について評価音声の前半及び後半の EMR を示している．ただし，縦軸は評価データの前半のカテゴリ感情，横軸は評価データの後半のカテゴリ感情とする．図より Prop. I〜III の EMR は Conv. の EMR よりも全体的に値が高いことが分かる．特に，感情ペアに “neutral” を含む評価データでは，音素クラス属性の数を増やす程 EMR が改善する可能性が示された．一方，前半が “anger” で後半が “sadness” の評価データについては，使用した方法やモデルに関わらず一貫して EMR が低い．感情ペアに “neutral” を含む評価データは，その他の 3 つの感情（“anger”，“hapiness”，“sadness”）からなる感情ペアの評価データよりも感情を推定しやすい可能性がある．また，前半が “anger” で後半が “sadness” の評価データについては，感情ペアの順序が逆の場合よりも EMR が高くなった．これは，感情が表出する順序によってフレーム単位の感情認識の難易度が変化する可能性を示している．以上の結果から，提案手法の結果は従来手法よりも高い EMR を示しており，提案手法はフレーム単位の音声感情認識の性能向上に有効であることが示された．

図 4.7 は，フレーム単位の評価結果が全体的に高かった wav2vec2+FC を各手法で学習した認識器の出力例を示している．評価用音声には “neutral” から “happiness” に感情が変化する音声を利用した．尚，図 4.7a はフレーム単位の評価用音声のメルスペクトログラム，図 4.7b〜4.7e は各手法で学習したモデルの出力（対数尤度）を示している．また，図 4.7a の青点線と図 4.7b〜4.7e の橙点線は正解感情ラベル列の感情が切り替わる地点を示している．故に，点線以前が “neutral” の音声，点線以降が “happiness” の音声となっている．

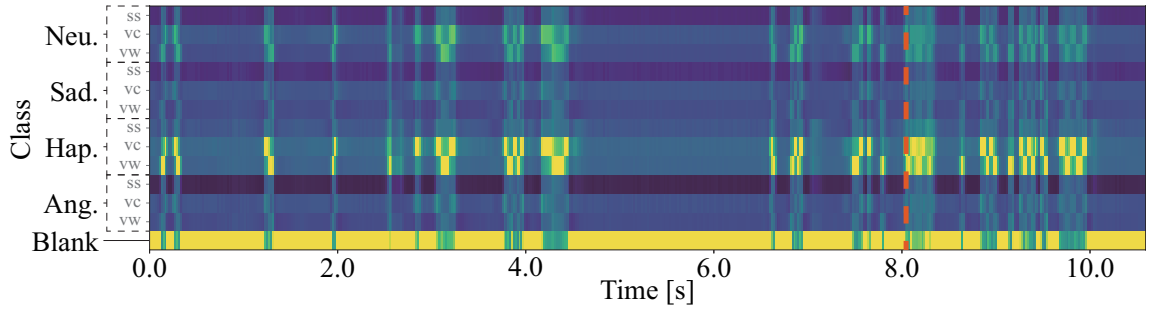
図 4.7b〜4.7e を比較すると，いずれの場合も音声が発話されている区間に対して感情ラベル列を予測し，発話されていない区間には blank 記号が予測されている．このことから，発声区間に着目して感情を推定できる認識器が学習されていると言える．また，考慮する音素クラス属性が増えるほど点線前後で高い対数尤度を示す感情クラスが変化していることが分かる．特に，図 4.7e では点線前後で対数尤度が高い感情クラスの傾向が “neutral” から “happiness” に変わっていることが分かる．このことから，各音素の音響的な違いを考慮することで従来手法よりも細かく正確な感情認識が可能になったと言える．



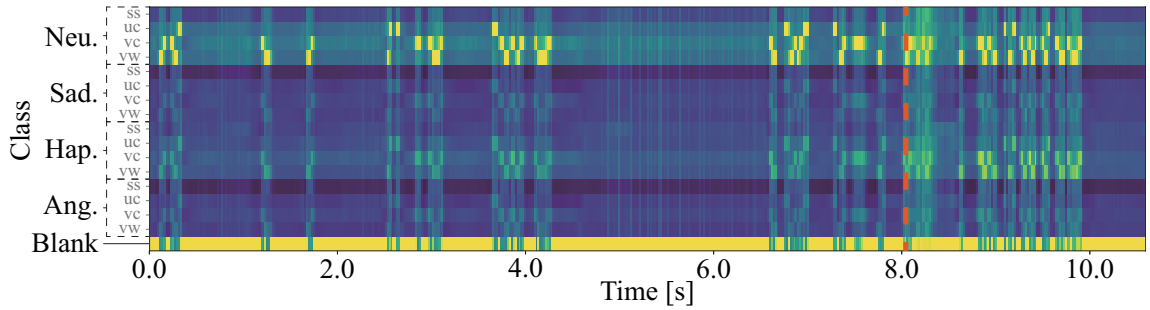
(a) メルスペクトログラム



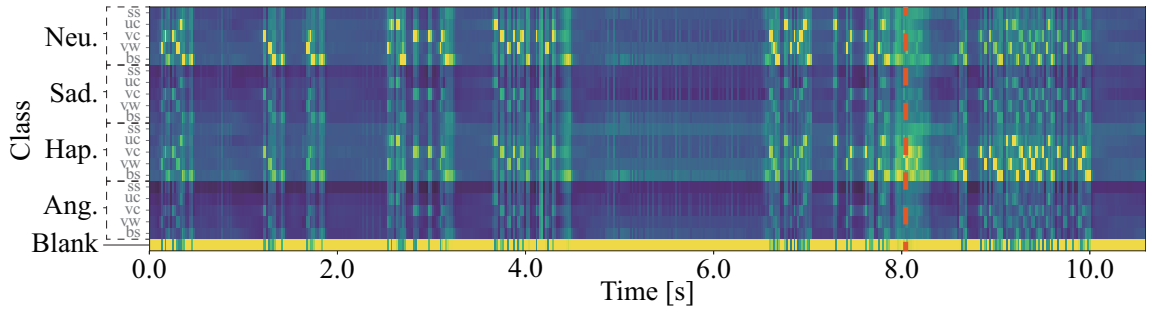
(b) Conv.



(c) Prop. I



(d) Prop. II



(e) Prop. III

図 4.7: 従来手法及び提案手法で学習した認識器の出力例

4.6 まとめ

本研究では、従来研究で考慮されていなかった音素クラス属性を持つ感情ラベル列を用いたフレーム単位の音声感情認識の学習手法を提案した．結果より，提案手法は発話単位とフレーム単位の両方でフレーム単位の音声感情認識の精度を向上することが分かった．特に，Prop. II は従来手法に比べて WA を最大 2.4%ポイント，UA を最大 8.2%ポイント向上させることができた．また，Prop. III は wav2vec2.0+FC を用いた場合で，EMR を従来手法よりも最大 2.4%ポイント向上させることができた．加えて，母音と有声・無声子音，特殊記号を考慮することで，モデルが発話内の細かな感情変化を認識できることが分かった．提案手法で学習したフレーム単位の音声感情認識は，発話内で変化する感情をより正しく認識できる．

第5章　感情キャプションを活用した音声感情認識

5.1　はじめに

既存の音声感情認識では，予め用意されたカテゴリや次元軸を用いて推定した感情を表現する．そのため，既知の代表的なカテゴリや次元軸では示せないような具体的な感情の認識には限界がある．例えば,「勝利に興奮し満足感を感じている」のような感情は喜怒哀楽やポジティブ－ネガティブだけで表現することが難しい．

この課題を解決し，感情軸に制限されない音声感情認識を実現するために，本章では音声が伝える感情の説明文（感情キャプション）を音声感情認識の予測結果として扱う手法の検討に取り組む．これらの技術を実現することで技術は，自身の心理状態分析やカウンセリングの補助，聴覚障がい者や外国語学習者のための感情表現字幕などへの応用が期待される．

本章では，次の 2 つの感情キャプションを活用した音声感情認識の研究に取り組んだ．一つ目は音声から感情キャプションを自動で書き起こす音声感情キャプショニングの研究である．既存の感情音声に感情キャプションを付与し，自動で感情キャプションを書き起こす手法の検討を行った．二つ目は推論時に定義した感情キャプションに基づくゼロショット音声感情認識である．学習時に定義されていない感情であっても，推論時に感情キャプションで自由に感情を定義できる手法の検討を行った．

5.2　音声感情キャプショニング

音声感情キャプショニングとは，入力音声から感情キャプションを推定することである．この技術の課題の一つに利用できるデータが限られている点が挙げられる．感情キャプションの推定モデルの構築には，既知のカテゴリや次元軸だけでは表現できない感情を示すキャプションと音声が対になったデータが必要である．特に，説明の自由度が高い感情キャプションを音声に対して人手で付与するには，多くの時間が必要になるため，効率的に集められる手段を考える必要がある．別の課題として，既存手法のように小規模データを用いた実験では，予測感情キャプションに含まれる語彙数が少ない点が挙げられる．そのため，感情キャプション以外の大規模なデータで事前に学習されたモデルを活用する必要がある．

故に，本研究では音声感情キャプショニングのためのデータ作成方法やモデル構築方法について検討する．はじめに，既存の感情音声データを利用し，GPT4 とクラウドソーシングを活用して感情キャプションを付与する．人が感情キャプションを記述する作業を GPT4 に置き換えるこ

とで，収集の効率化を期待する．また，収集した感情キャプションと人が記述した感情キャプションを比較し，どの程度妥当な感情キャプションが得られるか調査する．収集した感情キャプションは入力音声から感情キャプションを書き起こすモデルの構築に利用する．感情キャプションの生成に様々な LLM (large language model) を利用し，その結果を比較する．

5.2.1 問題の定式化とモデルの学習方法

問題の定式化

式 (5.1) に音声感情キャプショニングの算出方法を示す．ただし，入力系列長が K の音声特徴量を $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K]$，出力系列長が L の単語 ID 系列を $\mathbf{y} = [y_1, \ldots, y_L]$ とする．また，任意の $\mathbf{X}$ が入力された場合，ある単語 ID 系列 $\mathbf{y}$ が予測される確率を $P(\mathbf{y}|\mathbf{X})$，$P(\mathbf{y}|\mathbf{X})$ が最大になるときの単語 ID 系列を $\hat{\mathbf{y}}$ とする．

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{X}) \tag{5.1}$$

$P(\mathbf{y}|\mathbf{X})$ の導出には様々な手段が考えられるが，本研究ではニューラルネットワークを用いた深層学習手法で算出する．

Seq2seq による音声感情キャプショニング

Fig. 5.1 に seq2seq による音声感情キャプショニングの概要を示す．本研究で利用した seq2seq はエンコーダとデコーダ，注意機構で構成されている．エンコーダでは，入力の音声特徴量から潜在的な情報を抽出する．エンコーダでの処理を式 (5.2) に示す．ただし，系列長が K の音声特徴量 $\mathbf{X}$ をエンコーダ Enc に入力した場合に得られる出力ベクトル系列を $\mathbf{H}^{\mathrm{enc}} = [\mathbf{h}_1^{\mathrm{enc}}, \ldots, \mathbf{h}_K^{\mathrm{enc}}]$，時刻 k の出力ベクトルを $\mathbf{h}_k^{\mathrm{enc}}$ とする．

$$\mathbf{H}^{\mathrm{enc}} = \mathrm{Enc}(\mathbf{X}) \tag{5.2}$$

デコーダでは，エンコーダの出力ベクトルと一つ前の予測単語 ID から現在の予測単語 ID を推定する．このとき，注意機構でエンコーダの出力ベクトルの重み付け和を算出し，文脈ベクトルとしてデコーダで利用する．文脈ベクトルの算出とデコーダでの処理をそれぞれ式 (5.3), (5.4) に示す．ただし，l 番目の単語予測に利用する文脈ベクトルを $\mathbf{c}_l$，$\mathbf{c}_l$ の算出時にエンコーダから得られる k 番目の出力ベクトルに乗算する注意重みを $\alpha_{l,k}$，$\mathbf{c}_l$ と一つ前の y_{l-1} をデコーダ Dec に入力した場合に得られる出力ベクトル系列を $\mathbf{H}^{\mathrm{dec}} = [\mathbf{h}_1^{\mathrm{dec}}, \ldots, \mathbf{h}_L^{\mathrm{dec}}]$，$l$ 番目の出力ベクトルを $\mathbf{h}_l^{\mathrm{dec}}$ とする．

$$\mathbf{c}_l = \sum_{k=1}^{K} \alpha_{l,k} \mathbf{h}_k^{\mathrm{enc}} \tag{5.3}$$

$$\mathbf{h}_l^{\mathrm{dec}} = \mathrm{Dec}(\mathbf{c}_l, y_{l-1}) \tag{5.4}$$

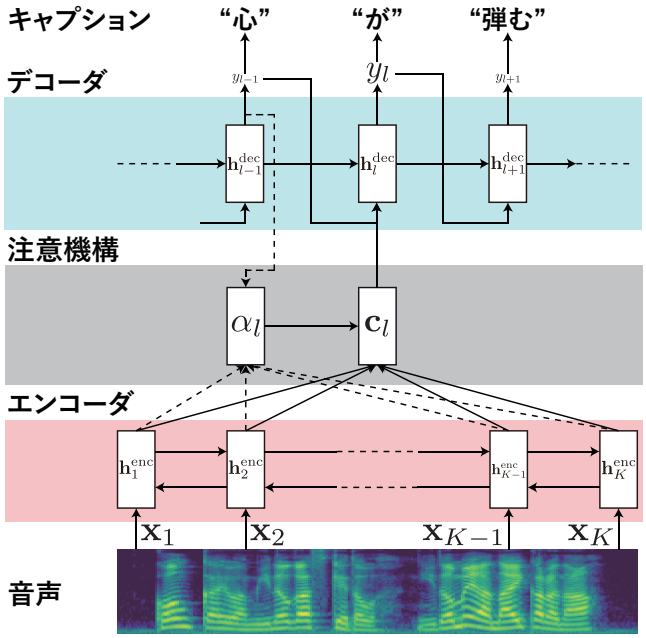


図 5.1: Seq2seq を用いた音声感情キャプショニングの概要

$\mathbf{H}^{\mathrm{dec}}$ から算出した予測単語 ID 系列の確率と正解の単語 ID 系列との交差エントロピー損失 $\mathcal{L}_{\mathrm{sac}}$ を最小化し，感情キャプションの推定モデルを学習する．

LLM を活用した音声感情キャプショニング

近年，GPT4 や Llama などの生成系 LLM（large language model）を活用した音声理解の研究が盛んに取り組まれている．例えば，Xu らは音声埋め込み表現を調整することで，指示文に従って LLM が感情の説明文を出力できるようにする手法を提案している [65]．また，Ando らは音声埋め込み表現と LLM を調整することで，指示文に従って LLM が話し方の説明文を出力できるようにする手法を提案している [66]．故に本研究でも LLM を感情キャプションのデコーダとして活用する手法を検討する．図 5.2 に，LLM を活用した音声感情キャプショニング手法の概要を示す．エンコーダでは，第 5.2.1 節の式 (5.2) と同様に，入力の音声特徴量から音声埋め込み表現を抽出する．尚，エンコーダの各層から得られる音声埋め込み表現に，学習可能な重みを付けて和を算出する．その後，音声埋め込み表現を Q-Former (querying-Transformer) に入力し，固定系列長の埋め込み表現を取得する．Q-Former は音声や画像などの埋め込み表現と LLM の埋め込み表現の間の差を埋めるためのモデルである [67]．Q-Former の自己注意機構と交差注意機構の処理をそれぞれ式 (5.5)，(5.6) に示す．ただし，系列長 N の学習可能な query を $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_N]$，$n$ 番目の学習可能な query を $\mathbf{q}_n$，自己注意機構及び交差注意機構の query，key，value に乗算する学習可能な重みをそれぞれ，$\mathbf{W}_{\mathrm{Q}}^{\mathrm{self}}$, $\mathbf{W}_{\mathrm{K}}^{\mathrm{self}}$，$\mathbf{W}_{\mathrm{V}}^{\mathrm{self}}$，$\mathbf{W}_{\mathrm{Q}}^{\mathrm{cross}}$, $\mathbf{W}_{\mathrm{K}}^{\mathrm{cross}}$，$\mathbf{W}_{\mathrm{V}}^{\mathrm{cross}}$，自己注意機構及び交差注意機構の key の次元数をそれぞれ $d_{\mathrm{K}}^{\mathrm{self}}$，$d_{\mathrm{K}}^{\mathrm{cross}}$ とする．また，自己注意機構及び交差注意機構の出力

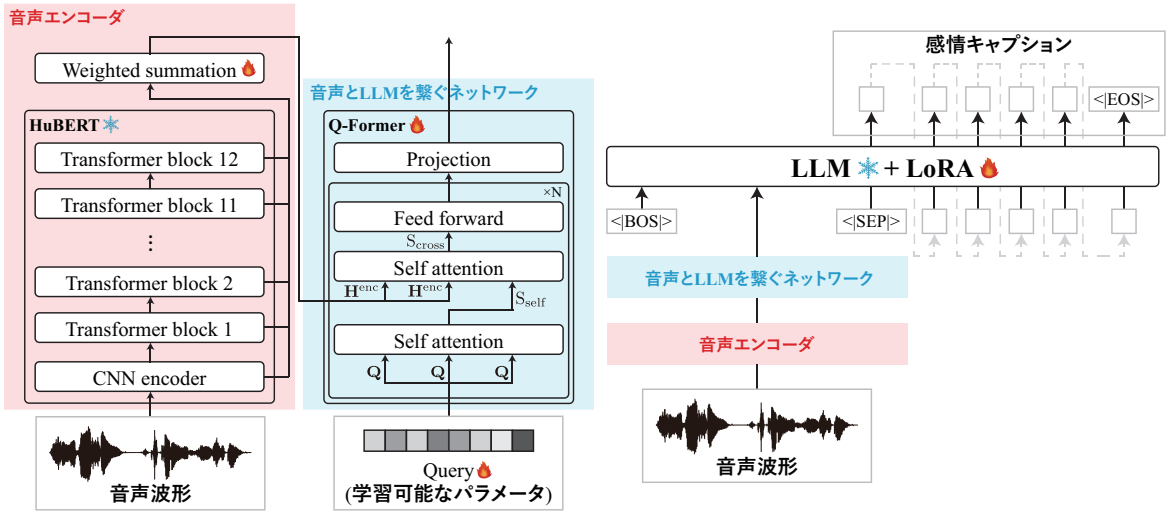


図 5.2: LLM を活用した音声感情キャプショニングの概要

をそれぞれ $\mathbf{S}_{\text{self}}$，$\mathbf{S}_{\text{cross}}$，ソフトマックス関数を softmax とする．

$$\mathbf{S}_{\text{self}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_{\text{Q}}^{\text{self}}(\mathbf{Q}\mathbf{W}_{\text{K}}^{\text{self}})^{\mathbb{T}}}{\sqrt{d_{\text{K}}^{\text{self}}}}\right)\mathbf{Q}\mathbf{W}_{\text{V}}^{\text{self}} \tag{5.5}$$

$$\mathbf{S}_{\text{cross}} = \text{softmax}\left(\frac{\mathbf{S}_{\text{self}}(\mathbf{H}^{\text{enc}}\mathbf{W}_{\text{K}}^{\text{cross}})^{\mathbb{T}}}{\sqrt{d_{\text{K}}^{\text{cross}}}}\right)\mathbf{H}^{\text{enc}}\mathbf{W}_{\text{V}}^{\text{cross}} \tag{5.6}$$

Q-Former を提案した論文中 [67] には，事前に多モーダル間の対照学習を行なう過程があるが，本研究では従来研究 [65] に合わせて実施しない．Q-Former から得られた固定系列長の埋め込み表現は，LLM の埋め込み表現としてデコーダに入力する．尚，デコーダに使用する LLM は LoRA(low-rank adaptation) で再調整する．LoRA とは事前学習済み LLM のモデルパラメータを固定したまま，各層に挿入した低ランク行列の値を更新することである [68]．これにより，LLM 全体の数%のパラメータ更新だけで LLM の再調整が行うことが出来る．最終的に，LLM から得られた予測単語 ID 系列の確率と正解の単語 ID 系列との交差エントロピー損失 $\mathcal{L}_{\text{sec}}^{\text{llm}}$ を最小化し，感情キャプションの推定モデルを学習する．

5.2.2 GPT4 とクラウドソーシングを用いた感情キャプションの収集とその評価

本章では，GPT4 とクラウドソーシングを用いて既存の感情音声データに感情キャプションを付与した．また，GPT4 とクラウドソーシングを用いて収集した感情キャプションを人が記述した感情キャプションと比較し，妥当性を評価した．

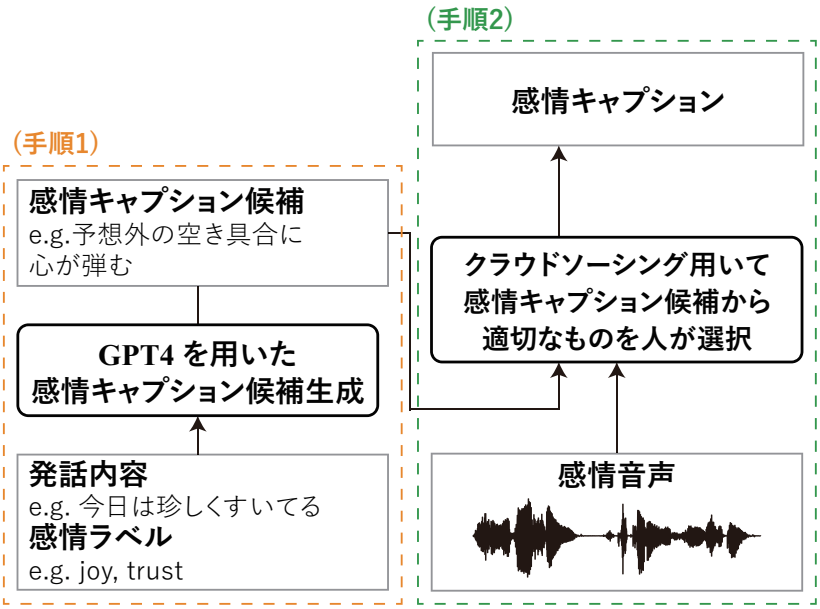


図 5.3: 感情キャプション付与手順の概要

音声感情キャプションの収集手順

既存の感情音声データセットとして JTES (Japanese twitter based emotional speech) [44] を利用した．JTES には，4 感情（喜び，悲しみ，怒り，平静）毎に 50 文ずつ用意した発話文について男女各 50 名ずつが演技発話した音声が収録されている．これらの音声に対応した日本語の感情キャプションを付与した．本研究で行った音声への感情キャプション付与の手順は図 5.3 及び次に示す通りである．

手順 1 発話文と正解感情から感情キャプションの候補を作成

手順 2 候補から音声に対して適切な感情キャプションを選択

手順 1 では，GPT4 (gpt-4-1106-preview) を利用して JTES に収録されている発話文とそれに対応する正解感情から感情キャプションの候補を作成した．GPT4 とは OpenAI が提供する LLM の一つであり，入力指示文の内容に適した回答を生成できる．このような LLM はこれまでにもデータ拡張 [69] や感情データセット構築 [70] の研究などに活用されている．本研究においても感情キャプション候補を人手でではなく GPT4 で生成する．

GPT4 で感情キャプション候補を作成するための指示文のテンプレートを図 5.4 に示す．図中の例を GPT4 に入力すると [**output**] に「怒りと嫌悪感を抱きつつ，あきれている」のような感情キャプション候補が生成される．指示文が英語である理由は，日本語よりも英語の指示文の方が適切に文書生成できる可能性が高いためである [71]．指示文には「提示されたテキストとカテゴリ感情から筆者の感情を説明せよ」という内容と回答例，GPT4 に与えるテキストとカテゴリ感情を記載した．また，JTES に含まれる「喜び」や「怒り」などの正解感情だけでなく様々な単語で説明された感情キャプションを生成するために，GPT4 に提示するカテゴリ感情の種類を表 5.1

テンプレート: 感情キャプション候補を生成するための指示文

Given one sentence and emotions, please describe
the writer's state of mind specifically and concisely in Japanese.
EXAMPLE 1: (× N)
e.g.
Sentence: おかしいって思わなかったの
Emotion: anger
State of mind: 些細なことが気になってイライラしている
INPUT AND OUTPUT:
Sentence: e.g. いつもふざけてばっかりだね
Emotion: e.g. anger, disgust
State of mind: [**output**]

図 5.4: GPT4 に与える指示文のテンプレート

表 5.1: GPT4 に与えたカテゴリ感情

Ground truth	Emotions given with prompts
anger	{anger.}, {anger, neutral}, {anger, anticipation}, {anger, disgust}, {anger, anticipation, neutral}, {anger, disgust, neutral}, {anger, anticipation, disgust}, {anger, anticipation, disgust, neutral}
joy	{joy}, {joy, neutral}, {joy, trust}, {joy, anticipation}, {joy, trust, neutral}, {joy, anticipation, neutral}, {joy, trust, anticipation}, {joy, trust, anticipation, neutral}
sadness	{sadness}, {sadness, neutral}, {sadness, disgust}, {sadness, surprise}, {sadness, disgust, neutral}, {sadness, surprise, neutral}, {sadness, disgust, surprise}, {sadness, disgust, surprise, neutral}
neutral	{neutral}, {neutral, anger}, {neutral, joy}, {neutral, sadness}, {neutral, anticipation}, {neutral, disgust}, {neutral, trust}, {neutral, surprise}

の通り追加した．正解感情が「平静」以外の 3 感情の場合，まず Plutchik の感情の輪に基づいて各正解感情とそれに隣接するカテゴリ感情,「平静」の感情を選択し，全ての組み合わせを考えた．その内，正解感情を含む組み合わせを指示文に利用した．正解感情が「平静」の場合，Plutchik の感情の輪に基づいて平静と基本 8 感情の組み合わせを考え,「平静」が含まれる組み合わせを指示文に利用した．感情キャプション候補は発話文につき重複無しで 100 キャプション以上になるよう GPT で生成した．

手順 2 では，手順 1 で作成した候補の内から感情音声に対して最も適当な感情キャプションを人手で選択した．本作業はクラウドソーシングにて一音声につき一感情キャプションが選ばれる

ように実施した．被験者には感情音声とそれに対応する感情キャプション候補の内，無作為に選んだ 4 つの文を提示した．その後,「音声を聞いた後に，4 つの選択肢の内から音声が伝える感情を最も適当に説明した文を一つ選んでください」のように指示した．以降，上記手順で選ばれた選択肢を感情キャプション（Semi-auto）と呼ぶ．

妥当性の評価方法

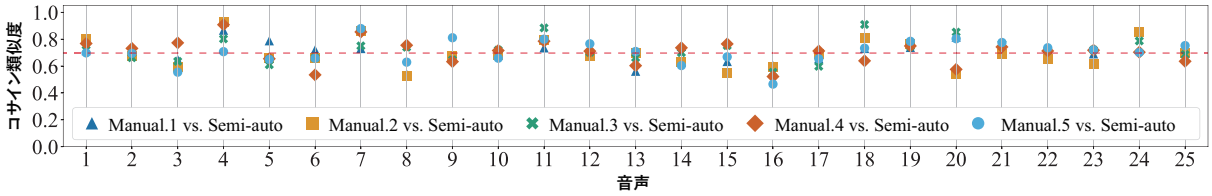
GPT4 とクラウドソーシングを用いて収集した感情キャプションの妥当性は，人が記述した感情キャプションと比較し評価する．はじめに，JTES から無作為に選んだ音声について人が記述した感情キャプションをクラウドソーシングで収集した．以降，人が記述した感情キャプションを感情キャプション（Manual）と呼ぶ．対象音声は 25 音声で，1 音声当たり聴取者は 5 名とした．

その後，感情キャプション（Semi-auto）と感情キャプション（Manual）を比較し，妥当性を客観的及び主観的に評価した．客観評価では，各感情キャプションの文埋め込み (`cl-nagoya/sup-simcse-ja-base`) について算出されたコサイン類似度を用いる．感情キャプション（Semi-auto）と感情キャプション（Manual）の類似度が全体の平均よりも高いペアが多い場合，感情キャプション（Semi-auto）は人が記述したものと同等に妥当であるとする．主観評価では，クラウドソーシングを用いて妥当性についての AB テストを行った．聴取者には音声と発話内容を提示し，感情キャプション（Semi-auto）と感情キャプション（Manual）のどちらが妥当であるか選択してもらった．対象の音声-感情キャプション対は 125 対で，1 対当たり聴取者は 5 名とした．感情キャプション（Manual）よりも感情キャプション（Semi-auto）の方が妥当であると選ばれた回数が同等以上の場合，感情キャプション（Semi-auto）は人が記述したものと同等に妥当であるとする．

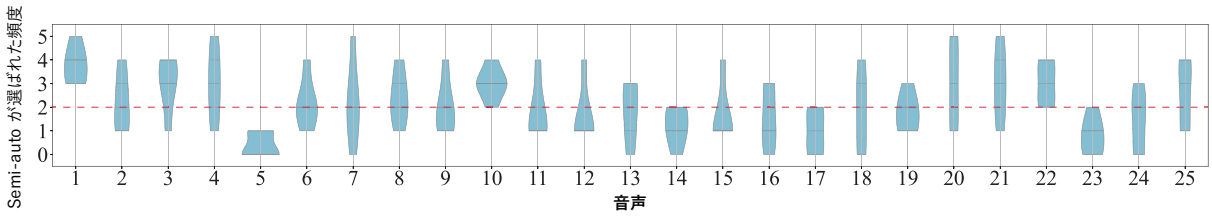
妥当性の評価結果

本研究で収集した感情キャプション（Semi-auto）と感情キャプション（Manual）についての発話毎の比較結果を図 5.5 に示す．表 5.5 (a) は各発話 1〜25 について，人が記述した 5 文の感情キャプション（Manual.1〜Manual.5）と感情キャプション（Semi-auto）の文埋め込みのコサイン類似度による客観評価を示している．表 5.5 (b) は各発話 1〜25 について，各感情キャプション（Manual）よりも感情キャプション（Semi-auto）の方が妥当であると選ばれた件数による主観評価を示している．

客観評価のコサイン類似度の平均は 0.706 となった．この平均よりも類似度が高くなったペアは全体の 51%となっており，感情キャプション（Semi-auto）はおおよそ妥当であると言える．コサイン類似度のばらつきを見ると，発話毎に傾向が異なることが分かる．最もコサイン類似度のばらつきが大きかった音声は 20 番「惜しかったね」であった．尚，20 番の音声が付与されている感情キャプション（Semi-auto）は「惜しい結果に驚きながらも落胆している」である．感情キャプション（Semi-auto）との類似度が最大の感情キャプション（Manual）は「結果を聞き心から残念な気持ちでいっぱいだ」となっており，説明している感情がほぼ一致していた．一方で，感情



(a) 客観評価



(b) 主観評価

図 5.5: 感情キャプション（Semi-auto）と感情キャプション（Manual）の比較

キャプション（Semi-auto）との類似度が最小になった感情キャプション（Manual）は「若干の煽りと嘲笑の感情が入り混じった発言」となっており，説明している感情が異なっていた．これは，話者や発話内容によって，聴取者が感じる感情がばらつきやすい音声がある可能性を示している．

主観評価値の平均は 2.09 となった．この平均よりも主観評価値が高くなったペアは全体の 59% となっており，主観評価においても感情キャプション（Semi-auto）はおおよそ妥当であると言える．主観評価値のばらつきを見ると客観評価と同様に発話毎によってばらついている．最も評価値のばらつきが大きい音声は 7 番「三時間って長すぎだろ」であった．尚，7 番の音声が付与されている感情キャプション（Semi-auto）は「不満に感じており、待ち時間に対する怒りと不快感を抱えている」である．感情キャプション（Semi-auto）の方が妥当であると評価者全員が選んだときの感情キャプション（Manual）は「長い待ち時間を聞いて率直に不快さを表している」であった．GPT4 の方が「不満」「怒り」「不快感」など複数の感情を説明していたため，妥当であると判断された可能性がある．また，感情キャプション（Manual）の方が妥当であると評価者全員が選んだときの感情キャプションは「予想外の長さに怒りを抑えきれない」であった．これは「予想外の長さに」と状況も説明しているため，具体的であり妥当であると判断された可能性がある．いずれの場合も感情キャプション（Manual）はネガティブな感情を示しており，明らかに意味合いが異なるキャプションではない．このことから，感情キャプション（Semi-auto）と感情キャプション（Manual）のどちらが妥当か判断が難しいことが分かる．

客観評価と主観評価を比較すると，類似度が低い場合であっても主観評価の値が高くなる場合がある．具体例として 4 番「三時間って長すぎだろ」の音声を取り上げる．4 番の発話には，感情キャプション（Semi-auto）に「長い待ち時間に強い不満を感じている」，感情キャプション（Manual）に「あまりに費やしたくない時間が長くてあきれている」が付与されている．これらの類似度は 4 番に付与されたキャプションの中では最も低い（0.708）が主観評価は 5 となっている．これは，感情キャプション（Semi-auto）にも感情キャプション（Manual）より妥当だと判断される文が含

まれることを示している．

以上より，GPT が生成した候補から人が選ぶ手順で収集した感情キャプション（Semi-auto）は，感情キャプション（Manual）と同等の妥当性を示すことが分かった．

5.2.3 音声感情キャプショニングのモデル構築と評価

本章では，第 3 章で収集した感情キャプション付き感情音声データを用いて音声感情キャプショニングのモデルを構築し，その性能を評価する．

学習データ

データセットには GPT とクラウドソーシングを用いて収集した感情キャプションを付与した JTES を用いる．尚，発話 ID が 1〜40 かつ 40 名の音声（12,800 文）を学習データ，発話 ID が 40〜45 かつ 10 名の音声（200 文）を検証データ，発話 ID が 45〜50 かつ 10 名の音声（200 文）を評価データとした．モデルの学習時には，感情キャプションについて句読点や空白の削除，英字の半角大文字化などの正規化を行った．正規化した感情キャプションはデコーダの出力形式に合わせてトークン化される．分割した感情キャプションの前後には開始記号と終端記号を追加した．

モデル学習

モデルには Seq2seq モデルと LLM を活用したモデルを用意した．各モデルで用いたエンコーダとデコーダについて表 5.2 に示す．

Seq2seq モデルのエンコーダには 3 層の BLSTM，デコーダには 1 層の LSTM，注意機構には local-aware attentioin [72] を利用した．モデルの入力は 80 次元の対数メルスペクトログラム，出力は予測した単語 ID 系列である．

LLM を活用したモデルのエンコーダには，日本語音声対応の事前学習済み HuBERT (`rinna/japanese-hubert-base`) を利用した [73]．Q-Former には従来研究 [65] を参考に BERT のモデル構造を利用し，日本語事前学習済みモデルパラメータ (`cl-tohoku/bert-base-japanese-v3`) を初期値として利用した．デコーダには Llama (large language model Meta AI) [74] を用いた事前学習済み LLM を利用した．本実験では日本語データを用いて一から学習した LLM-ja-3-3.7B (`llm-jp/llm-jp-3-3.7b`) と Sarashina2-7B (`sbintuitions/sarashina2-7b`)，事前学習済み Llama に追加で日本語データを学習した Llama-3-Swallow-8B (`tokyotech-llm/Llama-3-Swallow-8B-v0.1`) と Llama-3-youko-8B (`rinna/llama-3-youko-8b`) を用いた．いずれも Hugging Face [49] にて事前学習済みモデルパラメータが共有されている．モデルの入力は音声信号，出力は予測した単語 ID 系列である．

Seq2seq モデルの学習時には，エポック数は 50，バッチサイズは 16，学習率は 0.0001 とし，最適化手法には Adam を用いた．一方で LLM を活用したモデルの学習時には，エポック数は 20，

表 5.2: 各モデルのエンコーダ・デコーダ

	Model	
	Encoder	Decoder
(1)	BLSTM	LSTM
(2)	HuBERT+Q-Former	LLM-jp-3-3.7B
(3)		Sarashina2-7B
(4)		Llama-3-youko-8B
(5)		Llama-3-Swallow-8B

表 5.3: 予測感情キャプションの評価結果

	BLUE				METEOR	ROUGE_L	CIDEr	JaSPICE
	1	2	3	4				
(1)	0.291	0.167	0.090	0.052	0.167	0.259	0.135	0.054
(2)	0.273	0.158	0.076	0.041	0.159	0.247	0.144	0.064
(3)	0.268	0.164	0.091	0.056	0.163	0.246	0.176	0.046
(4)	0.279	0.171	0.099	0.062	0.174	0.260	**0.227**	**0.065**
(5)	**0.294**	**0.179**	**0.099**	**0.059**	**0.182**	**0.267**	0.189	0.060

バッチサイズは 8，勾配蓄積は 2，学習率は 0.0001 とし，最適化手法には AdamW(8bit) を用いた．

評価指標

音声感情キャプショニングのモデル評価には，画像キャプショニング [75] や環境音キャプショニング [76] の評価で良く用いられる指標である BLUE (bilingual evaluation understudy) [77]，ROUGE_L (recall-oriented understudy for gisting evaluation) [78]，METEOR (metric for evaluation of translation with explicit ordering) [79]，CIDEr (consensus-based image description evaluation) [80]，JaSPICE [81] を用いた．BLUE は正解及び予測キャプションの n-gram の適合率を算出する．ROUGE_L は正解及び予測キャプションで最も長い共通部分列について再現率重視の F 値を算出する．METEOR は正解及び予測キャプションのチャンクと unigram の一致数に基づく重み付けがされた再現率重視の F 値を算出する．尚，METEOR は日本語に非対応の為，本研究では言語非依存の METEOR を算出している．CIDEr は TF-IDF で重みづけた正解及び予測キャプションの n-gram コサイン類似度の平均を算出する．JaSPICE は日本語のシーングラフに基づいて値を算出する．これらの評価指標は，値が高くなるほど予測キャプションに含まれる単語や単語の並びが正解キャプションに類似していることを示している．モデルの予測キャプションについては，評価データのキャプションを正解として上記の評価値を算出する．

評価結果

Seq2seq と事前学習済み LLM として利用したモデル別に音声感情キャプショニング結果について表 5.3 に示す．尚，(1) は seq2seq モデルの出力結果，(2) から (5) はデコーダに事前学習済み LLM を用いたモデルの出力結果についての評価結果を示している．

(1) と (2) から (5) の結果を比較すると一部の結果を除いて事前学習済み LLM を使用したモデルの方が seq2seq を用いた場合よりも評価値が高くなった．特に Llama に追加で日本語データを学習させた LLM を用いた時が改善幅が大きかった．デコーダに LLM を用いることで事前学習済みの知識を活用可能になり，より評価データに含まれる単語を適切に含む予測が可能になったと考えられる．BLUE，METOR，ROUGE_L について最も評価値が高くなった手法は (5) となった．一方，CIDEr と JaSPICE について評価値が高くなった手法は (4) となった．これは，(4) と (5) では Llama3 に対して日本語データを追加で学習したモデルをデコーダに利用しており，事前学習時に利用した学習データ量が音声感情キャプショニング結果に影響していると考えられる．

表 5.4 には「怒り」「喜び」「悲しみ」「平静」の音声を入力とした音声感情キャプショニングの出力例を示している．尚，目標キャプションには学習データの正解信号として付与された感情キャプションを，予測キャプションには seq2seq を用いた (1) の出力結果と，llm を活用した手法の内，全体的に評価値が高かった (5) の出力結果を示す．

表 5.4a では (5) の結果のみ「怒り」の感情を説明している．また，(5) の結果には「無分別な行動」という単語は含まれていないが「相手の行動」が含まれており，(1) よりも具体的な感情を説明できていると言える．

表 5.4b では目標キャプションに含まれる「安心」と「喜び」という単語は (1) 及び (5) の結果に含まれてはいないが，いずれの結果も「喜び」の感情を説明している．特に (5) の結果には，発話内容に含まれる「間に合う」という単語が含まれている．これは，発話内容を加味して LLM が感情キャプションを予想しようとしている可能性を示している．

表 5.4c では (5) の結果のみ「悲しみ」の感情を説明している．また,「不平を言いつつも諦めている」説明と (5) の説明は意味合いも類似しており，比較的正確な感情キャプションの予測になっていると言える．

表 5.4d では目標キャプションに含まれる「平静」という単語は (1) 及び (5) の結果に含まれていないが，いずれの結果も「平静」の感情を説明している．また (1) 及び (5) の結果は「期待」の感情についても説明をしている．特に (5) の結果には「ワクワク」という単語が含まれていることから，目標キャプションにより近いキャプションを推定できている．

以上より，音声とそれに対する感情キャプションを収集すれば音声から感情キャプションをある程度予測できることが分かった．特に，デコーダに LLM を活用することでより具体的に感情を推定できることが分かった．

表 5.4: 音声感情キャプショニングの出力例

(a) 感情ラベルが「怒り」の音声

発話内容		むやみやたらと突っ込むな
目標キャプション		無分別な行動に対して怒りを感じている
予測キャプション	(1)	普通だが少し嫌悪感を感じている
	(5)	怒りと冷静さの中で相手の行動に対して批判的である

(b) 感情ラベルが「喜び」の音声

発話内容		間に合いそう
目標キャプション		安心して喜んでおり冷静さも保っている
予測キャプション	(1)	楽しさを伝えて相手に対する信頼と期待感を示している
	(5)	ぎりぎりで間に合って安堵している

(c) 感情ラベルが「悲しみ」の音声

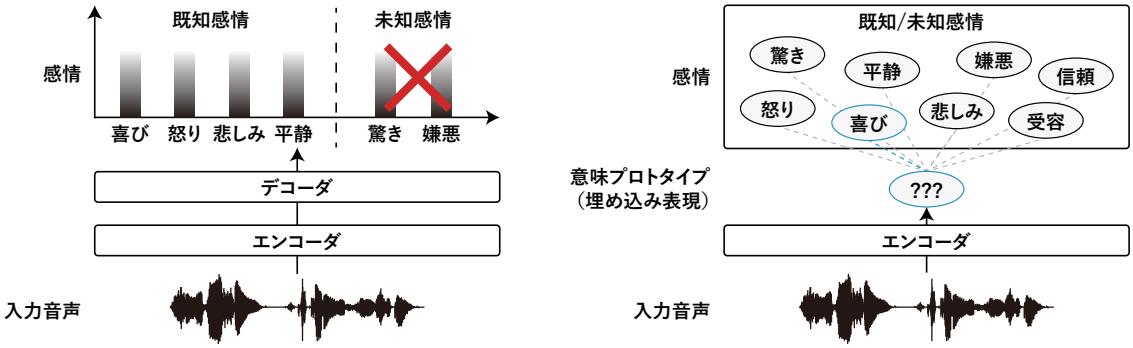
発話内容		相手が悪かったね
目標キャプション		運に不平を言いつつも諦めの気持ちが混ざっている
予測キャプション	(1)	冷静に事実を述べている
	(5)	少し残念に思いつつも、受け入れている

(d) 感情ラベルが「平静」の音声

発話内容		今日舞台の当落発表ですので、時間あるとき確認お願いします
目標キャプション		わくわくしながらも平静を保っている
予測キャプション	(1)	心穏やかでありながらも少し期待しているが冷静さを保っている
	(5)	普段通りに感じつつも、ちょっとしたワクワクを感じている

5.3 CLAP に基づくゼロショット音声感情認識を用いた購買意欲推定

多くの SER の研究では教師あり学習に基づいて認識モデルを学習することが多い．図 5.6a に教師あり学習に基づく音声感情認識の概要を示している．これらのモデルは事前に定義された既知感情は認識できるが，事前に定義されていない未知感情の認識は困難である．例えば，正解感情として学習時に定義された「怒り」や「喜び」などの認識は可能であるが，事前に定義されていない「驚き」や「嫌悪」などの感情は認識できない．この制限を取り払う手法として，学習時に定義されていない未知感情を推論できるゼロショット音声感情認識がある．図 5.6b にゼロショット音声感情認識の概要を示す．学習時には入力音声から感情の意味プロトタイプ（基本感情クラスや単語埋め込みなど）を推定するエンコーダを学習する．推論時には推論したい感情の意味プロトタイプを定義し，入力音声から得られる意味プロトタイプと最も類似する感情を最終的な認識結果とする．ゼロショット音声感情認識の従来研究では，音声特徴から感情属性を推定するモデ



(a) 教師あり学習に基づく音声感情認識　　**(b)** ゼロショット音声感情認識

図 5.6: 教師あり学習に基づく音声感情認識とゼロショット音声感情認識の概要

ルと感情属性から感情ラベルを推定するモデルを別々に学習し，入力音声を未知の感情クラスに分類する手法 [82] や，事前に定義されていないクラスの分類のために再構成された意味プロトタイプとデータ拡張を用いる手法 [83] などが提案されている．

ゼロショット音声感情認識の研究では，学習時に定義されていない未知の感情クラスを推論時にどう定義するかが課題である．従来手法では，単一の単語で定義できない感情クラスの推定は想定されていなかった．そのため，推論時に「購入したい気持ち（購買意欲）」のようにテキストで表現されたクラスに音声を分類することは困難である．様々な感情をゼロショットで認識可能にするためには，推論段階でクラスを自由に定義できる枠組みが必要である．また,「買いたい－買いたくない」のような二極性感情のゼロショット推定を実現したいという動機もある．

これらの要件を満たすために，本研究では双極性感情のゼロショット推定が可能な新しい CLAP（contrastive language-audio pretraining）手法を提案する．CLAP は推論時に分類クラスをテキストで表現できるため，感情クラスも同様に自由に決定できる．本手法では「快－不快」のような双極性サブクラスを複数定義し，多クラス感情に拡張した CLAP で分類モデルを訓練する．多クラス感情には，双極性サブクラスを持つ 6 つの基本感情を用いた．提案手法によって訓練されたモデルは，推論したいクラスが未知の双極性感情であっても 6 つの基本感情の知識を用いて感情を正しく分類できることが期待される．また，双極性感情としての購買意図に着目し，提案手法で訓練されたモデルが購買意図をゼロショットで推定できるか検証した．このような音声から直接購買意図を推定する手法の検討は本研究が初の取り組みである．結果より，提案したゼロショット音声感情認識の手法を場合でも教師あり学習モデルと同等の認識率で購買意欲を推定できることが分かった．

5.3.1 CLAP（Contrastive language-audio pretraining）

CLAP は，音響埋め込みとテキスト埋め込み間の類似性または非類似性を考慮したモデル学習手法である [84]．この手法で学習されたモデルは，推論時に分類クラスを単語で定義することができるため，音声を学習時に未定義のカテゴリに分類出来るようになる．また，単語だけでなく

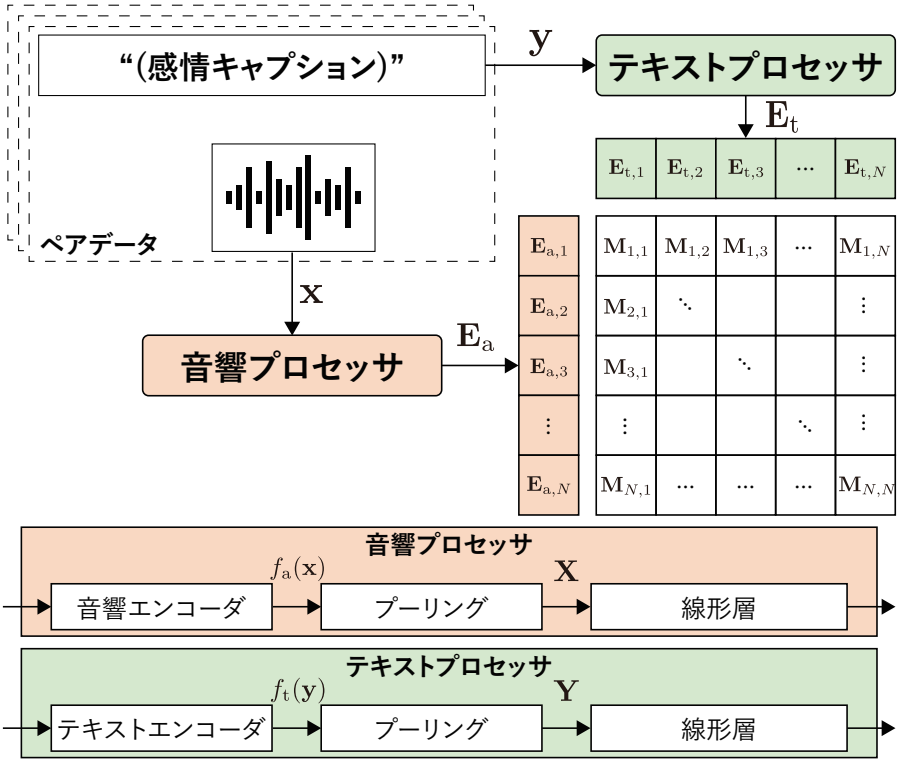


図 5.7: CLAP の概要（学習段階）

文でもカテゴリを定義できるため，モデルは自然言語で表現できる多様な感情カテゴリを推定できる．

図 5.7 及び 5.8 に学習段階と推論段階における CLAP モデルの処理を示す． 入力は音声信号と各カテゴリを示すテキストである．ここで，ミニバッチサイズが N のバッチの内，n 番目の音声信号とテキストをそれぞれ $\mathbf{x}_n = [x_1, \ldots, x_T]$, $\mathbf{y}_n = [y_1, \ldots, y_L]$ とする．また，それぞれ長さ T の音声信号と長さ L のテキストとする．

入力音声とテキストはそれぞれ音響エンコーダ及びテキストエンコーダで埋め込み表現に変換される．音響エンコーダ f_a とテキストエンコーダ f_t を，それぞれ式 (5.7) と式 (5.8) に示す．ただし，ミニバッチサイズが N のバッチの内，n 番目の音響埋め込み表現を $\mathbf{X}_n = [X_1, \ldots, X_{D_\mathrm{a}}]$，$n$ 番目のテキスト埋め込み表現を $\mathbf{Y} = [Y_1, \ldots, Y_{D_\mathrm{t}}]$，各埋め込み表現の次元数をそれぞれ D_a，D_t とする．また，Pooling は平均か選択による埋め込み表現のプーリングを示す．

$$\mathbf{X} = \mathrm{Pooling}(f_\mathrm{a}(\mathbf{x})) \tag{5.7}$$

$$\mathbf{Y} = \mathrm{Pooling}(f_\mathrm{t}(\mathbf{y})) \tag{5.8}$$

これらの埋め込み表現は式 (5.9) 及び (5.10) の通り，線形層に入力される．このとき，ミニバッチサイズが N のバッチの内，n 番目の線形層からの出力をそれぞれ $\mathbf{E}_\mathrm{a} = [E_{a,1}, \ldots, E_{a,D}]$ と $\mathbf{E}_\mathrm{t} = [E_{t,1}, \ldots, E_{t,D}]$，各出力の次元数を D とする．また，$\mathrm{Linear}_\mathrm{a}$ と $\mathrm{Linear}_\mathrm{t}$ はそれぞれ音響埋め込み

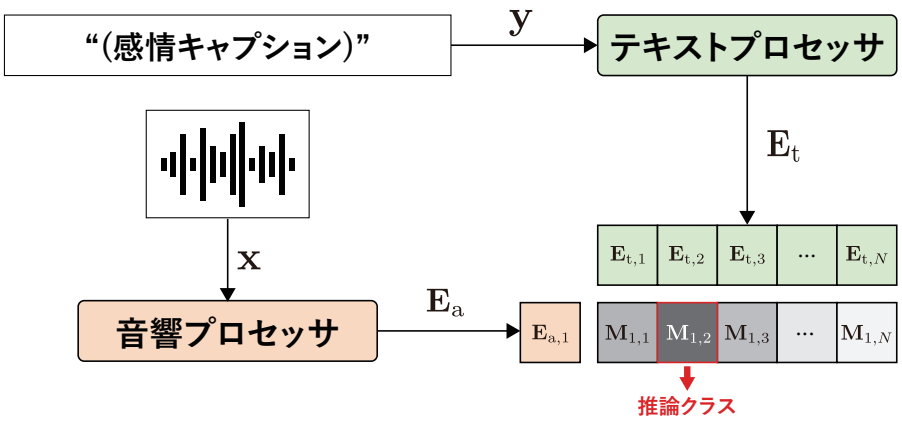


図 5.8: CLAP の概要（推論段階）

表現とテキスト埋め込み表現を入力する線形層とする．

$$\mathbf{E}_\mathrm{a} = \mathrm{Linear}_\mathrm{a}(\mathbf{X}) \tag{5.9}$$

$$\mathbf{E}_\mathrm{t} = \mathrm{Linear}_\mathrm{t}(\mathbf{Y}) \tag{5.10}$$

学習時は式 (5.11) に示す通り，$N \times N$ の $\mathbf{M}$ の類似度行列を算出する．尚，τ は温度パラメータである．

$$\mathbf{M} = \tau(\mathbf{E}_\mathrm{t} \cdot \mathbf{E}_\mathrm{a}^\top) \tag{5.11}$$

その後，式 (5.12) に示すシンメトリック損失関数 [85] で $\mathcal{L}$ を算出する．尚，CE はクロスエントロピー関数, $\hat{\mathbf{M}}$ は $N \times N$ の $\mathbf{M}$ 正解の類似度行列とする．

$$\mathcal{L}(\mathbf{M}, \hat{\mathbf{M}}) = \frac{1}{2}(\mathrm{CE}(\mathbf{M}, \hat{\mathbf{M}}) + \mathrm{CE}(\mathbf{M}^\top, \hat{\mathbf{M}})) \tag{5.12}$$

式 (5.12) を最小化することで，正解と不正解の区別をしながらモデルを学習する．

推定時は初めに分類カテゴリを単語または文で定義する．その後，音声及びテキストを各特徴に対応するエンコーダと線形層に入力し，固定長ベクトルを得る．最終的に各特徴から得られた固定長ベクトルを用いて音声とテキストの類似度を算出し，最も類似度が高いクラスを予測結果とする．

この手法は既存の音声感情認識にも活用されている．例えば，B. Elizalde らは CLAP を提案し，環境音データセットで訓練された CLAP モデルを用いてゼロショット音声感情認識の評価を行っている [84]．また，Y. Pan らは性別と感情分類のための GEmo-CLAP を提案し，学習時に既知の感情についての分類精度の向上を実現している [86]．しかし，感情音声データセットで訓練されたゼロショット音声感情認識を用いて，購買意図の有無などの双極性を持つ感情や，感情に関連するクラスを認識する研究は取り組まれていない．

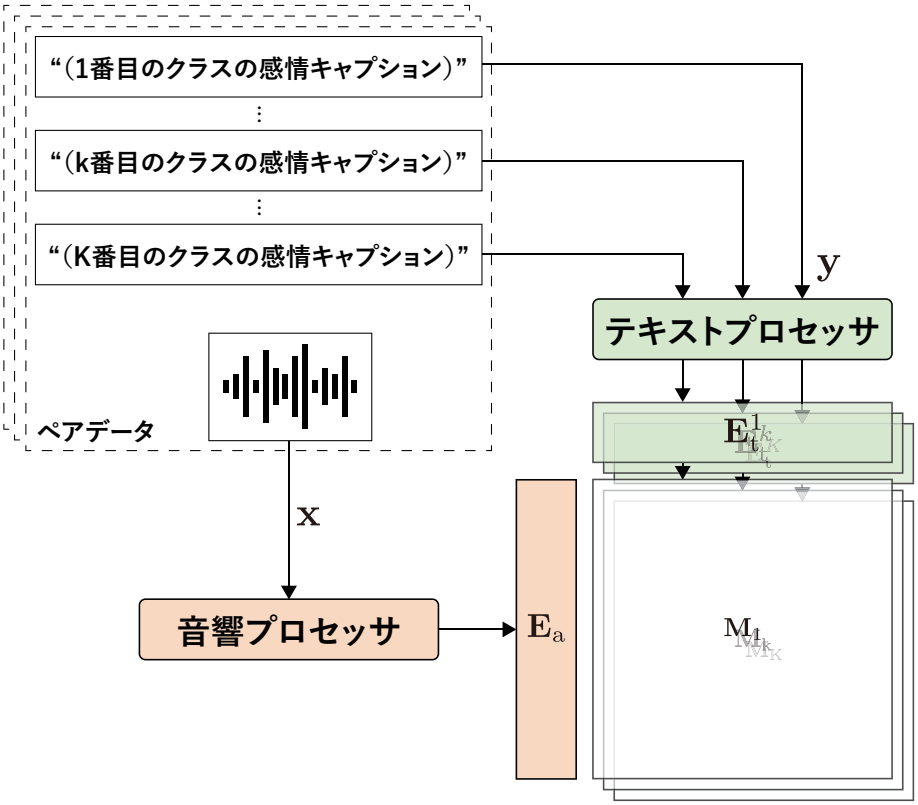


図 5.9: 提案手法の概要

5.3.2 多クラス－多タスク CLAP に基づく音声感情認識

本研究では，多クラス－多タスク CLAP と呼ばれるゼロショット音声感情認識の学習手法を提案する．提案手法の概要を図 5.9 に示す．

学習時，モデルは感情毎に用意されたマルチクラスの類似度行列を出力する．各感情の正解の類似度行列を用いてシンメトリック損失を計算した後，損失を合計してモデルを最適化する．損失 $\mathcal{L}_{\text{all}}$ の算出は式 (5.13) に示す通りである．K は感情クラスの数，k は感情クラスのインデックス，$\mathbf{M}_k$，$\hat{\mathbf{M}}_k$ は各感情における $N \times N$ の予測結果の類似度行列と $N \times N$ の正解の類似度行列を示している．

$$\mathcal{L}_{\text{all}} = \sum_{k=1}^{K} \mathcal{L}(\mathbf{M}_k, \hat{\mathbf{M}}_k) \tag{5.13}$$

推定時は従来の CLAP と同様に分類カテゴリを定義し，音声を分類する．提案手法で学習したモデルはサブクラスを持つ多クラス感情を認識できる．そのため，未知の双極性感情などをゼロショットで推定できることを期待した．

テンプレート: 言い換え処理の指示文

You are an imaginative assistant.
Given the emotion description, please generate ten paraphrase sentences in Japanese.
Note that you cannot add the prohibited word.

Input and output
Emotion description: e.g. 私の気持ちは覚醒しています
Prohibited word: e.g. 覚醒
Paraphrase sentences: [**output**]

図 5.10: 言い換え処理の指示文

5.3.3 言い換えによる多クラスデータ拡張

モデルの学習時に音声に対応する言語情報として感情キャプションを用意するが，それらの文の語彙や文法などの種類を増やすために，双極性を持つ感情のキャプションの言い換えによってデータ量を拡張した．言い換えには，OpenAI(`https://openai.com/gpt-4`) が提供している GPT4 を用いた．GPT4 に与えた指示文のテンプレートは表 5.10 に示す通りである．GPT4 は日本語よりも英語の指示文を与えられた方がより正確な回答を生成できるため [71]，本研究でも指示文を英語で記述した．指示文には双極性の感情を説明したキャプションと禁止単語を記述した．図中の例を GPT4 に入力すると [**output**] に「気持ちが澄み渡っていて鋭いです」のような入力感情キャプションを言い換えたテキストが生成される．尚，出力に禁止単語が含まれる場合は手動で修正した．

5.3.4 評価実験

本研究では音声から購買意思の有無についての二値分類を行い，提案手法の有効性を検証する．以降，実験に用いたデータセットと実験条件，実験結果について説明する．

データセット

提案手法の評価実験を行う前にモデル学習・評価のためのデータセットを構築した．初めにロールプレイングによる会話を収録し，短文毎に分割した．ロールプレイングでは，販売員と顧客との会話に関する 6 つのシナリオについて 5 人の日本語話者が自由に話した．尚，読み上げ用の文章は用意していない．録音時のサンプリングレートは 44,100 Hz であった．録音されたデータはシナリオ単位でセッションに分け，6 つのセッションからなるデータセットとした．その後，収集された音声に対して 6 つの感情（不快 − 快，眠気 − 覚醒，従順 − 支配，不信 − 信頼，無関心 − 関心，否定 − 肯定）の 7 段階の強度値（1. 非常に低い − 7. 非常に高い）を付与した．尚，聴取者は 1

表 5.5: 各感情の双極性サブクラスと各感情キャプションの対応

	負のラベル (0)	正のラベル (1)
不快 – 快	私は不快な気持ちです	私は快な気持ちです
睡眠 – 覚醒	私の気持ちは睡眠しています	私の気持ちは覚醒しています
服従 – 支配	私の気持ちは従属的です	私の気持ちは支配的です
不信 – 信頼	私は不信に思います	私は信頼しています
無関心 – 関心	私は無関心です	私は関心があります
否定 – 肯定	私は否定的です	私は肯定的です

音声につき 3 名になるようにした．これらの感情は関連研究 [87] に従って割り当てている．更に収集した発話の内，顧客の発言に限り 3 名の聴取者は購買意図を示す 7 段階の強度値を付与した．

最後に分類モデルの学習と評価のためにラベルを整理した．各聴取者が評価した強度値の分布は，全聴取者が付与した強度値の分布の平均と分散によって正規化された．正規化されたラベルについて最も頻度の高い強度値を閾値とし，閾値以下のスコアは負のラベル (0)，閾値以上のスコアは正のラベル (1) としてラベルを整理した．双極性サブクラスを持つ 6 感情の発話数は 2,598，双極的なクラスを持つ購買意図の発話数は 940 となった．尚，各発話のサンプリングレートは 16,000Hz に統一した．

提案手法のモデル学習と評価の正解信号には，上記で集約した双極性サブクラス基づく感情キャプションを用いた．表 5.5 は，各感情の双極性サブクラスと各感情キャプションの対応を示している．学習データについてデータ拡張を行わなかった場合と言い換えによるデータ拡張を行った場合の実験を行い，結果を比較した．評価では，6 つのセッションのうちの 1 つを評価用データとして，他セッションを学習用データとして使用した．

推定モデルの学習

提案手法のモデル構造を説明する．音響プロセッサは日本語事前学済み HuBERT [73] と線形層から構成される．HuBERT から得られる音声埋め込みと線形層から得られる中間特徴の次元数はそれぞれ 768 と 512 であった．音声埋め込みは線形層に渡す前に時間方向に平均化した．一方，テキストプロセッサは日本語事前学習済み DistilBERT [88] と線形層から構成される．DistilBERT から得られるテキスト埋め込みと線形層から得られる中間特徴の次元数もそれぞれ 768 と 512 であった．テキスト埋め込みの内，クラストークンのみを選択し線形層に渡す．HuBERT (`rinna/japanese-hubert-base`) と DistilBERT (`laboro-ai/distilbert-base-japanese`) の事前学習済みモデルパラメータには Hugging Face [49] が提供するものを利用した．

本論文では CLAP を利用せず購買意欲の有無を教師信号として学習した購買意図推定モデルをベースラインとした．ベースラインモデルの構造は HuBERT と線形層とし，HuBERT のモデルパラメータには提案方法と同じ事前学習済みのものを利用した．ベースラインにおける二項分類

表 5.6: 購買意欲の有無についての分類結果

手法	分類クラスを示す感情キャプション	WA	UA	再現率（購買意欲）	
				ない	ある
ランダム	-	49.5	49.4	48.8	50.0
教師あり学習 [ベースライン]	-	**74.0**	**69.2**	**78.5**	60.0
データ拡張なし					
ゼロショット [Ours]	(1) 私は買う気 { なし, あり } です	60.2	69.1	51.6	**86.7**
	(2) 私は購買意欲 { なし, あり }	64.2	60.5	67.7	53.3
	(3) 私は { 欲しくない, 欲しい } です	61.0	61.8	60.2	63.3
言い換えによるデータ拡張有					
ゼロショット [Ours]	(1) 私は買う気 { なし, あり } です	61.8	62.3	61.3	**63.3**
	(2) 私は購買意欲 { なし, あり }	65.0	63.3	66.7	60.0
	(3) 私は { 欲しくない, 欲しい } です	**73.2**	**69.8**	**76.3**	**63.3**

の閾値は，ROC (receiver operatorating characteristic) 曲線の Youden 指数によって定義される．エポック数は 300，バッチサイズは 64，学習速度は 0.000001，最適化方法は Adam [48] とした．ベースラインと提案方法の損失関数はそれぞれバイナリ交差エントロピー損失とシンメトリック損失であった．

評価指標

評価指標には WA (weighted accuracy) と UA (unweighted accuracy) を用いた．尚，WA は全データに対する正解率，UA は各クラスの正解率の平均である．また，各クラスの再現率についても比較する．これらは値が高いほどモデルがより正確に購買意図を推定できることを示している．

チャンスレートを示すためにランダムな整数値と正解信号の一致率を正解率として算出した．また，提案手法の評価には 3 通りの感情キャプションを用いた．これらの感情キャプションは全て購買意図を示しているが，文中で用いられている単語が異なっている．以上のチャンスレートとベースライン，ゼロショット推定の結果を比較する．

評価結果

表 5.6 は各手法の WA と UA，購買意欲の有無についての再現率を示している．チャンスレートは WA と UA の両方で約 50%となった．チャンスレートとベースラインの結果を比較すると，ベースラインの方が WA が 24.5%ポイント，UA が 19.8%ポイント値が高くなっていることが分

かった．これはベースラインのモデルがランダムな推定をしていないことを示している．また，深層学習を用いて音声から購買意図を予測するモデルを構築できることを示している．チャンスレートと提案手法の結果と比較すると，WA と UA は全ての提案手法でチャンスレートよりも高くなった．これは，提案手法によるゼロショット推定がランダムな推定をしておらず，学習時に未知の双極性感情を推定できるようにモデルが学習されていることを示している．

データ拡張を行わなかった場合の提案手法とベースラインの結果を比較すると，(1) の感情キャプションを用いた時の購買意図「ある」の再現率がベースラインよりも改善していることが分かる．また，(1) の感情キャプションを用いた時の UA と (3) の感情キャプションを用いたときの購買意図「はい」の再現率はベースラインと同等であった．これは，提案手法で学習したモデルは教師ありで学習したモデルと同程度のゼロショット推定が可能な場合があることを示している．一方，(1) の感情キャプションを用いたときの購買意図「いいえ」と (2) の感情キャプションを用いたときの購買意図「はい」の再現率はそれぞれ 60%以下であった．以上より，提案手法で学習したモデルはゼロショット推定時に用いる感情キャプションによって意味情報の対応関係を捉えられている場合とそうでない場合があることを示している．

データ拡張の有無で提案手法の結果を比較すると，データ拡張をした方が全体的に性能が高くなっている．特に，(3) の感情キャプションを用いた提案手法の結果は，教師ありで学習したベースラインの結果と同等である．これは，単語や文法などが異なる言い換えデータを学習データに追加したことで，音声と感情キャプションの意味的な対応関係を比較的容易に学習できるようになったことを示している．また，データ拡張よって音声ーテキストの対応関係の学習が進んだことにより，学習データに含まれていない「買う気」や「購買意欲」などの単語が含まれていない (3) のような感情キャプションを用いた時の認識性能が向上した可能性がある．

提案手法の確かな有効性を調べるために，ベースラインと (3) の感情キャプションを用いた提案手法の評価結果について有意差検定を行った．両者の間に有意差がない場合，ベースラインと提案手法で誤認識した回数は同じであると仮定し，両側符号検定を行った．結果，ベースラインと (3) の感情キャプションを用いた提案手法でデータ拡張を行わなかった場合の結果についての p 値 p は 0.05 未満であった ($p < 0.05$)．一方，ベースラインと (3) の感情キャプションを用いた提案手法でデータ拡張を行った場合の結果についての p 値 p は 0.05 以上であった ($p > 0.05$)．これら結果は，ベースラインと提案手法の結果に有意差が見られないことを示している．即ち，(3) の感情キャプションと言い換えによるデータ拡張を併用すると，提案手法は教師あり学習で学習した時と同等のゼロショット推定が行えることを示している．

5.4 まとめ

本章では，カテゴリ感情や次元感情よりも具体的に感情を説明した感情キャプションを活用した音声感情認識を紹介した．前半では，音声が伝える感情を説明した感情キャプションを自動で推定する音声感情キャプショニングについて説明した．感情音声に対する感情キャプションの収

集と分析，seq2seq や LLM を活用した音声感情キャプショニングの検討を行った．結果，GPT4 とクラウドソーシングを活用することで効率的に人が記述するような感情キャプションを感情音声に付与できる可能性を示した．また，収集したデータを用いて，音声からの感情キャプショニングを実現することができた．

後半では，感情キャプションを活用したゼロショット音声感情認識について説明した．学習時に定義されていない未知の双極性感情を認識するための多クラス－多タスク CLAP を提案した．結果，提案手法と言い換えによるデータ拡張を併用することで，教師あり学習モデルと同程度の精度でゼロショットな購買意図推定を実現することができた．

カテゴリ感情や次元感情よりも多様な感情を示すことができる感情キャプションを活用することで，音声感情キャプショニングや購買意欲なども推定できるゼロショット音声感情認識が可能になった．しかし，これらは感情キャプションの文脈などは考慮されていない．例えば「それは良い提案ですね。」と発言した後に「少し考えたいと思います。」と発言した場合,「提案に興味を惹かれ，期待感に満ち溢れている」のような感情が推測できるが，一方で「他の人も同じような提案をしています。」と発言した後に「少し考えたいと思います。」と発言した場合,「冷静に物事を判断しようと平静を保っている」のような感情が推測できる．このように，今後は文脈などを考慮した感情キャプショニングやゼロショット音声感情認識などを実現し，認識できる感情の幅を広げたいと考えている．

第6章　結論

本研究では，多様な感情の認識を目指し「音響情報と言語情報を併用した音声感情認識」「感情ラベル列を用いた音声感情認識」「感情キャプションを活用した音声感情認識」の手法を検討した．

第3章では，音響・言語情報の early fusion と late fusion を併用した音声感情認識について様々な融合処理を組み合わせた手法を同一条件下で比較した．結果，early fusion では和による融合，late fusion では結合による融合を利用した組み合わせ手法を用いた場合，従来手法の結果よりも約1.2%ポイント正解率が向上することが分かった．また，音響情報と音声認識結果を入力とする評価では平静の正解率が向上し，その他の感情の正解率が低下することが分かった．以上から，言語情報を併用することで，音響情報だけでは正しく認識されなかった音声を適切な感情に分類できるようになったと言える．

第4章では感情ラベル列を用いた音声感情認識について，音素クラス属性を取り入れた手法を提案した．結果，音素クラス属性付き感情ラベル列で学習したモデルによって，発話単位及びフレーム単位の感情の正解率を向上することができた．特に，母音と有声・無声子音を考慮した感情ラベル列を用いた場合，従来手法と比較して WA を最大2.4%ポイント，UA を最大8.2%ポイント向上させることができた．また，母音と有声・無声子音，基本記号を考慮した感情ラベル列を用いて wav2vec2.0+FC を学習した場合，EMR を従来手法よりも最大2.4%ポイント向上させることができた．以上から，細かな音素毎の音響的差異を考慮することで，時々刻々と変化する音声をフレーム単位で適切な感情に分類できるようになったと言える．

第5章では感情キャプションを活用した音声感情認識について，音声から自動で感情キャプションを推定する手法と，感情キャプションを活用したゼロショット音声感情認識の手法を提案した．音声感情キャプショニングについては，seq2seq モデルや LLM を活用したモデルを学習することで音声からの感情キャプションが可能であることを確認した．また，GPT4 とクラウドソーシングを活用することで効率的に人が記述するような感情キャプションを感情音声に付与できる可能性を示した．ゼロショット音声感情認識については，双極的なサブクラスを持つ感情に対応した CLAP を提案し，学習データにはない購買意欲の有無を教師あり学習モデルと同等の約70%の正答率で認識できる事を明らかにした．以上から，感情キャプションを用いることで，代表的な感情や事前に定義された感情に制限されない具体的な感情が認識できるようになったと言える．

本論文の研究を通して既存手法で認識できる感情の範囲を拡張し，多様な感情の認識を実現することが出来た．一方で，新たな課題も明らかになった．音響・言語情報を併用した音声感情認識では，音声認識をどのように感情認識モデルに統合するかが課題である．感情音声の単語誤り率を考慮した音声認識の構築や音声認識の潜在特徴を入力に利用した感情認識の検討などが必要で

ある．感情ラベル列を用いた音声感情認識では，フレーム単位の感情の認識率をどのようにして発話単位の感情の認識率に近づけるかが課題である．発話内で感情が変化するデータの収集及び作成とモデルの構築，より細かなフレーム単位のクラス属性を考慮した手法の検討などが必要である．感情キャプションを用いた音声感情認識では，入力音声の文脈の考慮が課題である．文脈が分かるような感情キャプション付き音声データの構築，事前学習済み LLM などが持つ学習データにない外部知識を活用した手法の検討が必要である．

謝辞

本研究を遂行するにあたり，6 年半に渡り不自由なく研究に専念できる環境を用意して頂いただけでなく，常に丁寧且つ熱心なご指導，ご鞭撻を頂いた立命館大学情報理工学部の 山下洋一 教授に深く感謝の意を表します．副査を勤めていただきました立命館大学情報理工学部 西浦敬信 教授，満上育久 教授には博士学位審査において多大なるご尽力を賜りました．深く感謝申し上げます．立命館大学情報理工学部の 福森隆寛 講師には，研究だけでなく後輩指導や進路について数多くのご助言を頂きました．深く感謝申し上げます．また，6 年半の間にお世話になりました，関西大学総合情報学部 山西良典 教授，同志社大学文化情報学部 井本桂右 准教授にも深く感謝申し上げます．その他，音声言語研究室の諸子には多くの研究の議論の機会を頂きました．心より感謝申し上げます．

株式会社 日立製作所研究開発グループ 住吉貴志 博士，川口洋平 博士，山下夏生 氏，土肥宏太 氏にはインターンでの研究と論文執筆にてお世話になりました．深く感謝申し上げます．また，6 年半の間にお世話になりましたインターン先の企業の皆様には，研究開発を学ぶ貴重な機会を頂きました．心より感謝申し上げます．最後に，経済面，精神面を支えてくださった家族と友人に感謝いたします．

参考文献

[1] H. Fujisaki, "Information, prosody, and modeling - with emphasis on tonal features of speech -," in *proceedings of Speech Prosody 2004*, Nara, Japan, Mar. 2004, pp. 1–10.

[2] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.

[3] Y. Gao, Z. Pan, H. Wang, and G. Chen, "Alexa, my love: Analyzing reviews of amazon echo," in *proceedings of 2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Guangzhou, China, Oct. 2018, pp. 372–380.

[4] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "AlloSat: A new call center French corpus for satisfaction and frustration analysis," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 1590–1597.

[5] M. Bojanič, V. Delič, and A. Karpov, "Call redistribution for a call center based on speech emotion recognition," *Applied Sciences*, vol. 10, no. 13, pp. 1–18, 2020.

[6] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in e-learning system based on affective computing," in *proceedings of Third International Conference on Natural Computation (ICNC 2007)*, vol. 5, Haikou, China, Aug. 2007, pp. 809–813.

[7] K.-Y. Huang, C.-H. Wu, M.-H. Su, and Y.-T. Kuo, "Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 393–404, 2020.

[8] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *proceedings of INTERSPEECH 2009 – 10^{th} Annual Conference of the International Speech Communication Association*, Brighton, United Kingdom, Sep. 2009, pp. 312–315.

[9] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *proceedings of INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, San Francisco, United State, Sep. 2016, pp. 2001–2005.

[10] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[11] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *proceedings of PlatCon 2017 – 2017 International Conference on Platform Technology and Service*, Busan, Korea, Feb. 2017, pp. 1–5.

[12] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of blstm neural networks for enhanced emotion classification in dyadic spoken interactions," in *proceedings of ICASSP 2012 – 2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 4157–4160.

[13] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based lstm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[14] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *proceedings of INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 2803–2807.

[15] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[16] Y. Chiba, T. Nose, and A. Ito, "Multi-stream attention-based blstm with feature segmentation for speech emotion recognition," in *proceedings of INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 3301–3305.

[17] J. Parry, E. DeMattos, A. Klementiev, A. Ind, D. Morse-Kopp, G. Clarke, and D. Palaz, "Speech emotion recognition in the wild using multi-task and adversarial learning," in

proceedings of INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, Sep. 2022, pp. 1158–1162.

[18] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *proceedings of INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czech, Sep. 2021, pp. 3400–3404.

[19] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *proceedings of INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czech, Sep. 2021, pp. 4508–4512.

[20] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition," in *proceedings of INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 3755–3759.

[21] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *proceedings of ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, May 2022, pp. 6922–6926.

[22] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *proceedings of ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, May 2022, pp. 6907–6911.

[23] Z. Zhao, Y. Wang, and Y. Wang, "Multi-level fusion of wav2vec 2.0 and BERT for multimodal emotion recognition," in *proceedings of INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 4725–4729.

[24] Y. Wu, Z. Zhang, P. Peng, Y. Zhao, and B. Qin, "Leveraging multi-modal interactions among the intermediate representations of deep transformers for emotion recognition," in *proceedings of MuSe'22 – 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, Lisboa, Portugal, Oct. 2022, pp. 101–109.

[25] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[26] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. Schuller, "Towards temporal modelling of categorical speech emotion recognition," in *proceedings of INTERSPEECH 2018 – 19th*

Annual Conference of the International Speech Communication Association, Hyderabad, India, Sep. 2018, pp. 932–936.

[27] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Frontiers in Psychology*, vol. 4, 2013.

[28] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[29] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[30] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," *Shape, Contour and Grouping in Computer Vision*, pp. 319–345, 1999.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of NIPS'12 – Advances in neural information processing systems 25*, vol. 25, Lake Tahoe, Nevada, United States, Dec. 2012, pp. 1097–1105.

[32] D. H. Hubel, and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[33] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[34] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, Dec. 1997.

[35] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *proceedings of ICLR'17 – the 5th International Conference on Learning Representations*, Toulon, France, Apr. 2017, pp. 24–26.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *proceedings of NIPS'17 – the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, Dec. 2017, pp. 6000–6010.

[37] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *proceedings of IEEE SLT 2018 – 2018 IEEE Spoken Language Technology Workshop*, Athens, Greece, Dec. 2018, pp. 112–118.

[38] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," in *proceedings of APSIPA ASC2019 – 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Lanzhou, China, Nov. 2019, pp. 519–523.

[39] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *proceedings of INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018, pp. 247–251.

[40] M. Chen, and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," in *proceedings of INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 374–378.

[41] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, May 2020, pp. 6484–6488.

[42] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *proceedings of EMNLP 2027 – 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sep. 2017, pp. 1103–1114.

[43] O. Verkholyak, A. Dvoynikova, and A. Karpov, "A bimodal approach for speech emotion recognition using audio and text," *Journal of Internet Services and Information Security*, vol. 11, pp. 80–96, 01 2021.

[44] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *proceedings of O-COCOSDA 2016 – 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques*, Bali, Indonesia, Oct. 2016, pp. 16–21.

[45] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *proceedings of INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018, pp. 2207–2211.

[46] K. Maekawa, "Corpus of spontaneous japanese: its design and evaluation," in *proceedings of SSPR 2023 – ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, Apr. 2003, p. paper MMO2.

[47] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *proceedings of INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, Aug. 2017, pp. 1089–1093.

[48] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," in *proceedings of ICLR'15 – 3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[49] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *proceedings of EMNLP 2020 – 2020 Conference on Empirical Methods in Natural Language Processing System Demonstrations*, Online, Oct. 2020, pp. 38–45.

[50] T. Kudo, "Mecab : Yet another part-of-speech and morphological analyzer," 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:61584143

[51] S. Nakagawa, and H. Takagi, "Statistical methods for comparing pattern recognition algorithms and comments on evaluating speech recognition performance," *Journal of the Acoustical Society of Japan*, vol. 50, no. 10, pp. 849–854, 1994.

[52] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *proceedings of ICSLP 2004 – 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, Oct. 2004, pp. 889–892.

[53] A. Aryani, M. Conrad, and A. Jacobs, "Extracting salient sublexical units from written texts: "emophon," a corpus-based approach to phonological iconicity," *Frontiers in Psychology*, vol. 4, 2013.

[54] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *proceedings of INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018, pp. 3688–3692.

[55] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. W. Schuller, "Attention-Enhanced Connectionist Temporal Classification for Discrete Speech Emotion Recognition," in *proceedings of INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 206–210.

[56] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, and B. W. Schuller, "Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition," *Neural Networks*, vol. 141, pp. 52–60, 2021.

[57] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *proceedings of ICML'06 – the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA, June 2006, pp. 369–376.

[58] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[59] H. Feng, S. Ueno, and T. Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr model," in *proceedings of INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 501–505.

[60] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "Ctc-segmentation of large corpora for german end-to-end speech recognition," in *proceedings of SPECOM2020 – Speech and Computer: 22nd International Conference*, St. Petersburg, Russia, Oct. 2020, pp. 267–278.

[61] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *proceedings of NIPS'20 – 34th International Conference on Neural Information Processing Systems*, no. 1044, Vancouver, BC, Canada, Dec. 2020, pp. 12 449–12 460.

[62] J. Kahn, M. Riviére, W. Zheng, E. Kharitonov, Q. Xu, P. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *proceedings of ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, May 2020, pp. 7669–7673.

[63] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *proceedings of ICASSP 2015 – 2015 IEEE International*

Conference on Acoustics, Speech and Signal Processing, South Brisbane, QLD, Australia, Apr. 2015, pp. 5206–5210.

[64] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *proceedings of ICLR'20 – 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, Apr. 2020.

[65] Y. Xu, H. Chen, J. Yu, Q. Huang, Z. Wu, S.-X. Zhang, G. Li, Y. Luo, and R. Gu, "SECap: Speech emotion captioning with large language model," in *proceedings of AAAI-24 – the 38th Annual AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, Vancouver, Canada, Feb. 2024, pp. 19 323–19 331.

[66] A. Ando, T. Moriya, S. Horiguchi, and R. Masumura, "Factor-conditioned speaking-style captioning," in *proceedings of INTERSPEECH 2024 – 25rd Annual Conference of the International Speech Communication Association*, Kos, Greece, Sep. 2024, pp. 782–786.

[67] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *proceedings of ICML'22 – the 39th International Conference on Machine Learning*, vol. 162, Baltimore, Maryland, USA, July 2022, pp. 12 888–12 900.

[68] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *proceedings of ICLR'22 – the 10th International Conference on Learning Representations*, Online, Apr. 2022.

[69] L. Fang, G. Lee, and X. Zhai, "Using GPT-4 to augment unbalanced data for automatic scoring," in *arXiv preprint, arXiv:2310.18365*, 2023.

[70] D. Xin, J. Jiang, S. Takamichi, Y. Saito, A. Aizawa, and H. Saruwatari, "Jvnv: A corpus of japanese emotional speech with verbal content and nonverbal expressions," *IEEE Access*, vol. 12, pp. 19 752–19 764, 2024.

[71] OpenAI, "GPT-4 technical report," in *arXiv preprint, arXiv:2303.08774*, 2023.

[72] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *proceedings of NIPS'15 – the 28th International Conference on Neural Information Processing Systems*, Montreal Canada, Dec. 2015, pp. 577–585.

[73] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. 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.

[74] Meta, "The llama 3 herd of models," in *arXiv preprint, arXiv:2407.21783*, 2024.

[75] L. Qiao, and W. Hu, "A survey of deep learning-based image caption," in *proceedings of CEI 2022 – the 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology*, Fuzhou, China, Sep. 2022, pp. 120–123.

[76] X. Mei, X. Liu, M. D. Plumbley, and W. Wang, "Automated audio captioning: an overview of recent progress and new challenges," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, 2022.

[77] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *proceedings of ACL'02 – the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318.

[78] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81.

[79] S. Banerjee, and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, June 2005, pp. 65–72.

[80] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *proceedings of CVPR 2015 – IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, Oct. 2015, pp. 4566–4575.

[81] Y. Wada, K. Kaneda, and K. Sugiura, "JaSPICE: Automatic evaluation metric using predicate-argument structures for image captioning models," in *proceedings of CoNLL 2023 – the 27th Conference on Computational Natural Language Learning*, Singapore, Dec. 2023, pp. 424–435.

[82] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. W. Schuller, "Autonomous emotion learning in speech: A view of zero-shot speech emotion recognition," in *proceedings of INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 949–953.

[83] X. Xu, J. Deng, Z. Zhang, Z. Yang, and B. W. Schuller, "Zero-shot speech emotion recognition using generative learning with reconstructed prototypes," in *proceedings of ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes, Greece, June 2023, pp. 1–5.

[84] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *proceedings of ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes, Greece, June 2023, pp. 1–5.

[85] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *proceedings of ICML'21 – the 38th International Conference on Machine Learning*, vol. 139, Online, July 2021, pp. 8748–8763.

[86] Y. Pan, Y. Hu, Y. Yang, W. Fei, J. Yao, H. Lu, L. Ma, and J. Zhao, "GEmo-CLAP: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition," in *proceedings of ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, Apr. 2024, pp. 10 021–10 025.

[87] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, 2011.

[88] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *arXiv.preprint, arXiv:1910.01108*, 2019.

研究実績（筆頭著者）

原著論文

J.1 **永瀬 亮太郎**，福森 隆寛，山下 洋一: “音響・言語情報の early fusion と late fusion を併用した音声感情認識，” 日本音響学会誌，80 巻，5 号，pp. 244–252．2024.

査読付き国際会議

I.1 **Ryotaro Nagase**, Takeshi Sumiyoshi, Natsuo Yamashita, Kota Dohi, and Yohei Kawaguchi: “Can We Estimate Purchase Intention Based on Zero-shot Speech Emotion Recognition, ” in proceeding of *APSIPA ASC 2024 - Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec. 2024.

I.2 **Ryotaro Nagase**, Takahiro Fukumori and Yoichi Yamashita: “Speech Emotion Recognition by Estimating Emotional Label Sequences with Phoneme Class Attribute, ” in proceeding of *INTERSPEECH2023 - 24th Annual Conference of the International Speech Communication Association*, Dublin, Ireland, pp. 4533–4537, Aug. 2023.

I.3 **Ryotaro Nagase**, Takahiro Fukumori and Yoichi Yamashita: “Speech Emotion Recognition Using Label Smoothing Based on Neutral and Anger Characteristics, ” in proceeding of *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech2022)*, Hybrid form (Physical in Osaka, Japan and virtual), Mar. 2022.

I.4 **Ryotaro Nagase**, Takahiro Fukumori and Yoichi Yamashita: “Speech Emotion Recognition with Fusion of Acoustic- and Linguistic-Feature-Based Decisions, ” in proceeding of *APSIPA ASC 2021 - Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 725 – 730, Hybrid form (Physical in Tokyo, Japan and virtual), Dec. 2021.

国内会議

D.1 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“音声感情キャプショニングのためのデータ作成とモ

デル構築の検討，” 音声言語シンポジウム，Dec. 2024.

D.2 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“音声感情キャプショニング：感情の説明文を書き起こす音声感情認識の初期検討，” 日本音響学会 2024 年春季研究発表会，pp. 847–850，Mar. 2024.

D.3 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“感情ラベル列推定の知識蒸留を用いた音声感情認識，” 日本音響学会 2023 年秋季研究発表会，pp. 1157–1160，Sep. 2023.

D.4 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“音素クラス属性を考慮した感情ラベル列の推定による音声感情認識，” 音学シンポジウム，vol. 123，no. 88，pp. 42–47，Jun，2023.

D.5 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“音声認識とのマルチタスク学習を用いた CTC モデルに基づく短区間音声感情認識，” 日本音響学会 2022 年秋季研究発表会，pp. 1175–1178，Sep. 2022.

D.6 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“音響情報と言語情報を利用した短区間の音声感情認識, ” 日本音響学会 2022 年春季研究発表会，pp. 1149–1152，Mar. 2022.

D.7 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“声から感情を読み解く技術 – 言語情報も利用した音声感情認識 – , ” 日本音響学会関西支部第 24 回関西支部若手研究者交流研究発表会，Dec. 2021.

D.8 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“「平静」と「怒り」の感情の特性を考慮した音声感情認識のための label smoothing，” 日本音響学会 2021 年秋季研究発表会，pp. 1131–1134，Sep. 2021.

D.9 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“音声特徴とテキスト特徴の協調利用によるマルチモーダル感情認識，” 音声言語シンポジウム，2020-SLP-134，no. 10，pp. 1–6，Nov. 2020.

D.10 **永瀬 亮太郎**，福森 隆寛，山下 洋一：“テキスト情報を利用した深層学習に基づく音声感情認識，” 日本音響学会 2021 年春季研究発表会，pp. 975–978，Mar. 2021.

D.11 **永瀬 亮太郎**，井本 桂右，山西 良典，山下 洋一：“朗読音声を用いたノンパラレル声質変換による変換音声の話者性と表現の評価，” 電子情報通信学会技術研究報告，vol. 119，no. 440，pp. 213–218，Mar. 2020.

受賞

H.1 2023 年度立命館大学大学院情報理工学研究科 優秀研究賞

H.2 2023 年第 7 回 IEEE SPS Tokyo Joint Chapter Student Award

H.3 2021 年度立命館大学大学院情報理工学研究科 研究奨励賞

H.4 日本音響学会関西支部第 24 回関西支部若手研究者交流研究発表会 奨励賞

H.5 2021 年度立命館大学大学院リサーチプロポーザルコンテスト 優秀賞

研究実績（共著）

査読付き国際会議

I.1 Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, **Ryotaro Nagase**, Takahiro Fukumori, and Yoichi Yamashita: ”Environmental Sound Synthesis from Vocal Imitations and Sound Event Labels,” in proceeding of *ICASSP2024 - 49 th IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, pp. 411–415, Apr. 2024.

I.2 Noriyuki Tonami, Keisuke Imoto, **Ryotaro Nagase**, Yuki Okamoto, Takahiro Fukumori, Yoichi Yamashita: “Sound Event Detection Guided by Semantic Contexts of Scenes, ” in proceeding of *ICASSP2022 - IEEE 47 th International Conference on Acoustics, Speech and Signal Processing*, Singapore, pp. 801–805, May. 2022.

国内会議

D.1 岡本 悠希，**永瀬 亮太郎**，岡本 南美，齋藤 佑樹，福森 隆寛，山下 洋一：“環境音に対する印象説明文データセットの構築と分析，” 日本音響学会 2024 年秋季研究発表会，pp. 339–342，Sep. 2024.

D.2 岡本 悠希，井本 桂右，高道 慎之介，**永瀬 亮太郎**，福森 隆寛，山下 洋一：“環境音の模倣音声を利用した環境音合成とデータセット構築，” 日本音響学会電気音響研究会/電子情報通信学会応用音響研究会，p. 22，2024.

D.3 岡本 悠希，井本 桂右，高道 慎之介，**永瀬 亮太郎**，福森 隆寛，山下 洋一：“環境音の模倣音声を用いた環境音合成の検討とデータセット構築，” IDR ユーザフォーラム，Dec. 2023.

D.4 岡本 悠希，井本 桂右，高道 慎之介，**永瀬 亮太郎**，福森 隆寛，山下 洋一：“Voice-to-foley: 環境音を模倣した音声を入力とする環境音合成，” 日本音響学会 2023 年秋季研究発表会，pp. 1071-1074，Sep. 2023.

D.5 福森 隆寛，**永瀬 亮太郎**，許 凱，山下 洋一：“マスクおよびフェイスシールド着用話者の音声コーパス構築，” 日本音響学会 2023 年秋季研究発表会，pp. 1161–1164，Sep. 2023.

D.6 大澤まゆ子，**永瀬 亮太郎**，福森 隆寛，山下 洋一：“話題情報を用いた音声感情認識，” 2022年度人工知能学会全国大会，講演番号：2I5-OS-9a-04，Jun. 2022.

D.7 砺波 紀之, 井本 桂右, **永瀬 亮太郎**, 岡本 悠希, 福森 隆寛, 山下 洋一：“事前定義されていないシーン情報を利用可能な音響イベント検出，” 日本音響学会 2022 年春季研究発表会，pp. 243–246，Mar. 2022.