BERT と PANNs を用いた文脈を考慮した効果音検索

奥田　萌莉†　　大島　裕明†,††

† 兵庫県立大学 大学院情報科学研究科　〒 650–0047 兵庫県神戸市中央区港島南町 7–1–28
†† 兵庫県立大学 社会情報科学部　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
E-mail: †ad21d034@gsis.u-hyogo.ac.jp, ††ohshima@ai.u-hyogo.ac.jp

あらまし　本研究では，小説の文に対して，その文の文脈を考慮した効果音を検索する問題に取り組む．効果音検索の入力は，小説の文とその中で指定された一語である．出力は効果音である．たとえば,「怒りの雨が降っていた」という文と，その中の語である「雨」が入力となる．この入力に対しては，雨が激しく降りしきる効果音が検索されることを目指す．効果音を検索するとき，単なるテキスト検索だけでは効果音を十分に絞り込むことができない場合が多い．なぜなら，大量の効果音の中には，異なる音に同じキーワードが付与されていることがよくあるためである．そこで，本研究では 2 つの課題に取り組む．まず，テキスト検索によって同じキーワードが付与されている効果音を検索する課題に取り組む．次に，それらの効果音の中から，小説の文が表す文脈に合う効果音を絞り込む課題に取り組む．特に，小説の文が表す感情を推定し，その感情が効果音と一致するかを基準に，効果音を検索する．そのために，まず，あらゆるテキストを感情分類するモデルを構築する．さらに，感情と効果音が一致するかを学習したモデルを構築する．これらのモデルを組み合わせることで，課題を実現する．

キーワード　効果音，情報検索，BERT

1　はじめに

効果音は，映画，ドラマ，アニメなどの映像コンテンツやラジオや朗読などの音声コンテンツにおいて，重要な役割を果たすものである．本研究では，小説の文に対して，その文の文脈を考慮した効果音を検索する問題に取り組む．

本研究の応用として，テキストコンテンツの音声読み上げにおいて，効果音を自動的に付与することなどが考えられる．たとえば，Kindle などの電子書籍では，自動読み上げを行うことが可能である．

自動音声読み上げに自動的に効果音が付与されると，より臨場感が得られるといった効果が期待される．たとえば「怒りの雨が降っていた.」という文に対して，雨が激しく降りしきる効果音が付与される．このような場合,「怒りの雨が降っていた.」という文から,「怒り」の感情を読み取ることができる.「怒り」の感情を表す効果音のひとつとして，雨が激しく降りしきる効果音が挙げられる．以上のように，小説の文が表す感情と関連のある効果音が鳴らされる．

私たちは YouTube Audio Library [1]などのサイトで効果音を検索することができる．このようなサイトから，自動読み上げに付与される効果音を検索する．

効果音を検索するとき，単なるテキスト検索だけでは効果音を十分に絞り込むことができない場合が多い．効果音には，様々なキーワードが付与されている．しかし，大量の効果音の中には，異なる音に同じキーワードが付与されていることがよくある．たとえば,「水」というキーワードが付与された効果音が数多く存在する．それらの中には，波が打ち寄せる効果音や水道水が流れる効果音など，多種多様な「水」に関する効果音が含まれている．そのため，テキスト検索だけでは「水」を表す効果音の中から適切な効果音を十分に絞り込むことができない．

そこで，本研究では，小説の文とその中で指定された一語を入力し，入力の文と語の文脈を考慮した効果音を検索する問題に取り組む．たとえば,「怒りの雨が降っていた.」という文と，その中の語である「雨」が入力となる．この入力が与えられたときには，雨が激しく降りしきる効果音が検索されることを目指す．提案手法では，まず，テキスト検索によって，同じキーワードが付与されている効果音を検索する課題に取り組む．次に，それらの効果音の中から，小説の文が表す文脈に一致する効果音を絞り込む課題に取り組む．特に，小説の文が表す感情を推定し，その感情が効果音と一致するかを基準に，効果音を検索する．

詳細については 4 節で説明するが，本研究の全体像を図 1 に示した．大きく以下の 2 段階で構成される．

- テキストの類似度計算
- 感情と効果音の一致度判定

まず，入力の文と語に対して，効果音に付与されたテキストとの類似度を計算した．提案手法に対する一つ目の比較手法として，この類似度が高い順に，効果音をランキングした．

次に，入力の文と語が表す感情に対して，効果音との一致度を判定した．二つ目の比較手法として，この一致度が高い順に，効果音をランキングした．

最後に，計算された類似度と一致度の積を計算し，この積が大きい順に効果音をランキングした．提案手法として，この積が高い順に，効果音をランキングした．

1：YouTube　Audio　Library：https://studio.youtube.com/channel/UCC8x8OvEIdfmafahBJPfxkw/music

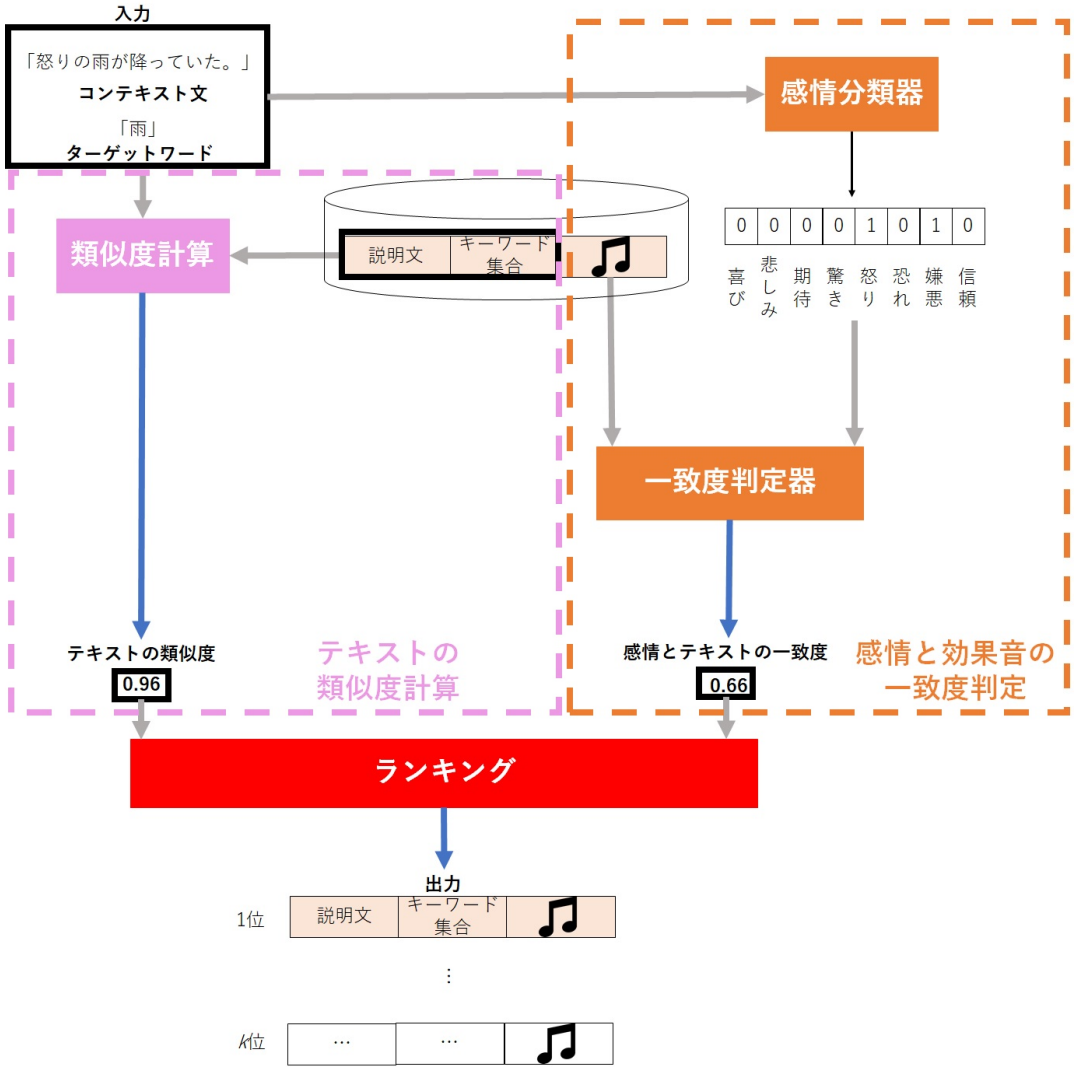

図 1　提案手法における全体像

2　関 連 研 究

本節では，オーデイオブックに関する研究，BERT，PANNs について紹介する．

Steinhaeusser ら [9] は，効果音と音楽を加えることによる影響に注目して，オーディオブックとロボットによるストーリーテリングを比較した調査を行った．結果，効果音を付けなかった場合のオーディオブックと比較して，効果音を付けた場合のロボットによるストーリーテリングの方が，人間の感情がより高まる傾向が示された．

Lopez ら [6] は，オーディオフィルムにおいて言語を最小限にすることは可能かどうかを追求した．効果音と音楽のみと，それに言語を加えたオーディオフィルムの 2 つを比較した．その結果，大多数のオーディオフィルムにおいて，最小限の言語と音だけで効果的に話が進んだと報告されている．

テキストをベクトル化するときには，BERT [2] がよく用いられる．BERT とは，自然言語処理におけるさまざまなタスクで利用されている深層学習のモデルである．BERT は自然言語処理を目的としたディープラーニングモデルの一つとして，2018 年に Google の Devlin ら [2] によって提案された．これは，Transformer [11] で構成されたモデルである．近年では日本語を事前学習させたモデルが公開され始めている．このような取り組みから，日本語における自然言語処理タスクにおいても BERT を活用する研究が増えてきている [12] [13].

音をベクトル化する手法として，PANNs [5]，wav2vec [1]，WaveRNN [3] ConvTasNet [8] などが知られている．これらのモデルの中で，wav2vec，WaveRNN，ConvTasNet といったモデルは，入力として自然言語の音声が対象である．それに対して，PANNs は，入力として効果音が対象である．PANNs とは，AudioSet と呼ばれる約 200 万の音データで事前に学習させた深層学習のモデルである．音のログメルスペクトラムを CNN に入れたベクトルと，音を CNN に入れたベクトルを結合して，音がどのような種類なのかを予測する．AudioSet データセットで PANNs をファインチューニングした結果，平均適合率が 43.9%であったと報告されている．

以上より，本研究では BERT モデルと PANNs モデルを用いて，テキストと音をベクトル化する．

表 1　SERIES 6000 THE GENERAL の 7,547 件における統計情報

	平均値	中央値	標準偏差
効果音	21.05 秒	6.03 秒	32.35 秒
説明文	23.87 文字	21 文字	12.30 文字
キーワード集合	1.66 語	2 語	0.61 語

3　問題定義と用いたデータ

本節では，本研究の問題定義と用いたデータについて述べる．

3.1　問 題 定 義

本研究における効果音検索の入力は，小説の文とその中で指定された語である．出力は，効果音である．

ここで，小説におけるテキストを**コンテキスト文**，そのテキストの中で効果音を付けたい語を**ターゲットワード**と呼ぶ．入力と出力の形は以下のようになる．

入力　コンテキスト文とターゲットワード

出力　効果音

本研究の目的は，コンテキスト文とターゲットワードを入力し，これらの入力の文脈を考慮した効果音を検索することである．たとえば,「怒りの雨が降っていた.」というコンテキスト文と「雨」というターゲットワードを入力し，1 位として雨が激しく降りしきる効果音，2 位として雨と雷の効果音を出力する．

「怒りの雨が降っていた.」というコンテキスト文を入力として与える．検索対象となる効果音には，たとえば「雨」というテキストが付与されているものが 32 件ある．これら 32 件には，雨が激しく降りしきる効果音，雨が降るなか雷が落ちる効果音，小雨が降る効果音などの様々な雨の音が含まれている．雨が激しく降りしきる効果音には,「水面に激しく雨が降る音.」という説明文が付与されている.「怒りの雨が降っていた.」というコンテキスト文の文脈は，この説明文の文脈と類似している．よって，テキスト同士の類似度を指標として効果音を検索できる．

「怒りの雨が降っていた.」というコンテキスト文の文脈を表す効果音は，多種多様であると考えられる．たとえば，雷が鳴る効果音，平手打ちする効果音，炎が燃え上がる効果音などが挙げられる．このような効果音には，コンテキスト文の文脈を表す共通点がある．よって，テキストの文脈と効果音の一致度を指標として効果音を検索できる．

3.2　効果音検索の対象となるデータ

効果音の検索対象として，SERIES 6000 THE GENERAL [2] を用いた．SERIES 6000 THE GENERAL には，7,547 件の効果音が含まれている．全データの音の長さを図 2 に示した．全データの約 59%が 10 秒以内である．

すべての効果音には，2 種類のテキスト情報が付与されている．たとえば，RAIN6.wav という効果音には,「雨」というキーワードと「水面に激しく雨が降る音.」という説明文が付与されている．それぞれを，**キーワード集合**，**説明文**と呼ぶこととする．全データの統計情報を表 1 に示した．各効果音の平均値は 21.05 秒，中央値は 6.03 秒，標準偏差は 32.35 秒である．説明文の平均値は 23.87 文字，中央値は 21 文字，標準偏差は 12.30 文字である．キーワード集合の平均値は 1.66 語，中央値は 2 語，標準偏差は 0.61 語である．すべての効果音に対して，説明文とキーワード集合が 1 つずつ対応している．

異なる効果音でも，説明文やキーワード集合が同じものが存在する．表 2 に，異なる効果音の中でキーワード集合が同じものの数を示した．キーワード集合が「自動車」である効果音の数は 513 と最も多く,「自動車, フォード」である効果音の数が 165 と次に多かった．表 3 に，異なる効果音の中で説明文が同じものの数を示した．説明文が「ガラスが勢い良く落ちて，ガシャンと砕ける音.」である効果音の数が最も多く,「ロボットのサーボ・モーター (間接調速装置) の作動音.」である効果音の数が次に多かった．表 4 に，異なる効果音の中でキーワード集合と説明文の両方が同じものの数を示した．説明文が「ガラスが勢い良く落ちて，ガシャンと砕ける音.」でキーワード集合が「ガラス，粉砕」ある効果音の数が最も多かった．

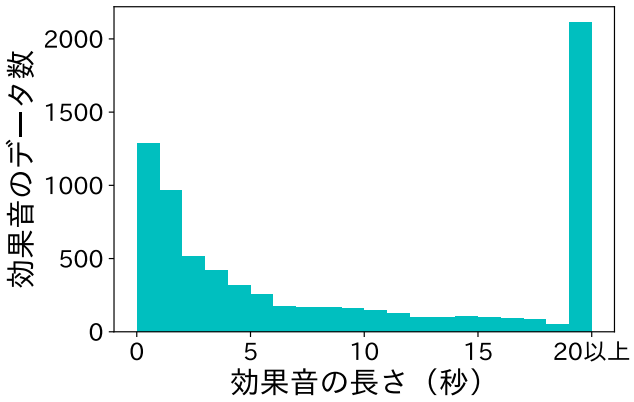

図 2　音のデータ数と音の長さのヒストグラム

表 2　キーワード集合が同じ効果音の数

キーワード集合	効果音の数
{ 自動車 }	513
{ 自動車, フォード }	165
{ 足音, 木床 }	155
{ 機関銃 }	129
{ 銃 }	87

表 3　説明文が同じ効果音の数

説明文	効果音の数
ガラスが勢い良く落ちて，ガシャンと砕ける音.	18
ロボットのサーボ・モーター (間接調速装置) の作動音.	15
大規模な爆発音.	14
大型で獰猛な動物が吠える声.	13
高電圧の火花.	13

2 : SERIES 6000 THE GENERAL : https://sonicwire.com/product/16681

表 4　キーワード集合と説明文の両方が同じ効果音の数

説明文	キーワード集合	効果音の数
ガラスが勢い良く落ちて，ガシャンと砕ける音.	ガラス, 破砕音	18
ロボットのサーボ・モーター (間接調速装置) の作動音.	ロボット, モーター	15
大規模な爆発音.	爆発	14
大型で獰猛な動物が吠える声.	動物, 吠え声	13
高電圧の火花.	電気, 火花	13

表 5　WRIME データセットの 4 段階のラベル数

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼	総数
0	29,883	32,006	29,369	32,348	41,771	34,086	35,679	41,282	276,424
1	7,682	7,610	8,963	8,135	886	6,986	5,672	1,733	47,667
2	4,512	3,167	3,970	2,284	366	1,808	1,457	171	17,735
3	1,123	417	898	433	177	320	392	14	3,774

4　効果音検索の構成

本研究では，コンテキスト文とターゲットワードを入力し，これらの入力の文脈を考慮した効果音を出力する．

提案手法では，大きく以下の 2 段階をとる．

- テキストの類似度計算
- 感情と効果音の一致度判定

テキストの類似度計算を行う理由，感情と効果音の一致度判定を行う理由について説明する．

まず，テキスト検索によって, 同じ説明文とキーワード集合が付与されている効果音を検索するために，テキストの類似度計算を行う．この計算により，コンテキスト文とターゲットワードに対して，テキスト情報における類似性をもつ説明文とキーワード集合を対応させることができる．

次に，同じ説明文とキーワード集合が付与されている効果音の中から，コンテキスト文とターゲットワードが表す感情に一致する効果音を絞り込むために，感情と効果音の一致度判定を行う．ここでの感情とは，コンテキスト文とターゲットワードが表す文脈のひとつである．たとえば,「怒りの雨が降っていた.」というコンテキスト文は「怒り」の感情を表すといえる．「怒り」の感情を表す効果音として，雷が鳴る音，平手打ちする音，炎が燃え上がる音などが考えられる．そのため，あらゆる効果音には，感情分類できる特徴があるといえる．

図 1 において，テキストの類似度計算と，感情と効果音の一致度判定の 2 段階についての概略を示した．

まず，図 1 のピンク色の部分に，テキストの類似度計算の概略を示した．コンテキスト文とターゲットワードに対して，音の説明文と音のキーワード集合との類似度を計算する．

次に，図 1 のオレンジ色の部分に，感情と効果音の一致度判定の概略を示した．この一致度判定は，コンテキスト文とターゲットワードが表す感情と，効果音との一致度を判定するものである．コンテキスト文とターゲットワードを感情分類し,「喜び」,「悲しみ」,「期待」,「驚き」,「怒り」,「恐れ」,「嫌悪」,「信頼」の 8 ラベルに対応するベクトルを得る．この感情とテキストの一致度を計算する．

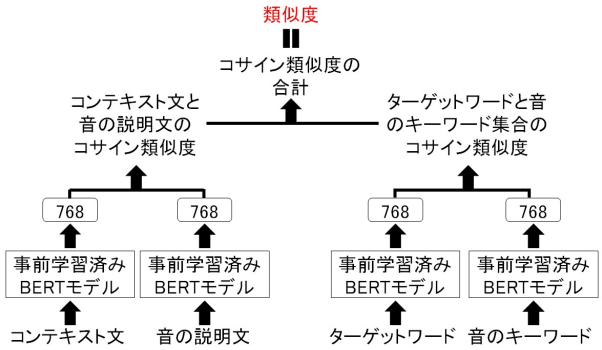

図 3　テキストの類似度計算

最後に，テキストの類似度と，感情とテキストの一致度の積を計算することで，Top k 件の効果音をランキングする．

4.1　テキストの類似度計算

本節では，同じ説明文とキーワードが付与されている効果音の検索について説明する．コンテキスト文とターゲットワードに対して説明文とキーワード集合との類似性を求めることで，効果音を検索する．計算された類似度の大きい順に，効果音をランキングする．コンテキスト文，ターゲットワード，説明文，キーワード集合をそれぞれベクトル化し，これらのベクトルのコサイン類似度を計算する．図 3 に，これらのコサイン類似度における計算方法を示した．

このモデルとして，東北大学の乾研究室が公開する BERT の事前学習モデルを用いる．利用した事前学習モデルは，日本語 Wikipedia を学習データとして，Masked Language Modeling タスクと Next Sentence Prediction タスクで事前学習されたものである．事前学習済みの BERT モデルに，ターゲットワードを入力し，入力長の 768 次元のベクトルが得られる．これらのベクトルを Average Pooling し，768 次元のベクトルを得る．同様に，コンテキスト文，キーワード集合，説明文をそれぞれ事前学習済みの BERT モデルに入力し，768 次元のベクトルを得る．

ターゲットワードとキーワード集合における 768 次元のベクトル同士のコサイン類似度を計算する．コンテキスト文と説明

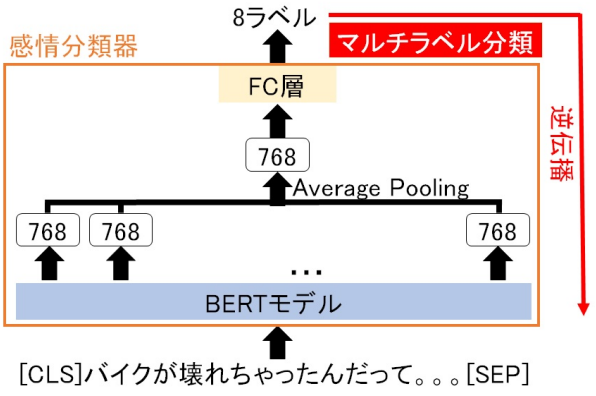

図 4　感情分類器の構造

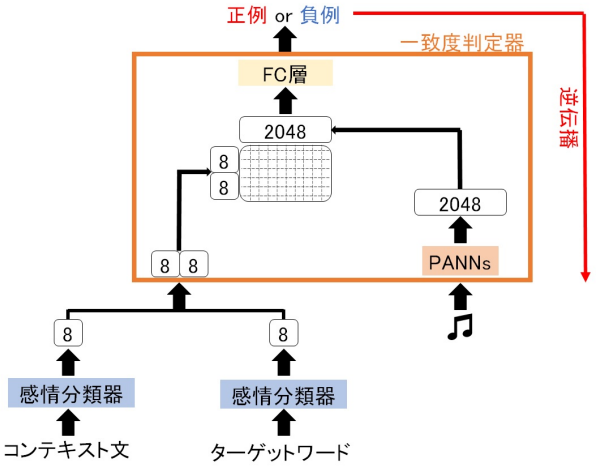

図 5　感情と効果音の一致度判定

文における 768 次元のベクトル同士のコサイン類似度を計算する．これら 2 つのコサイン類似度を合計し，効果音のランキングのための指標とした．

この指標の値が大きい順に，効果音のランキングを行った．

4.2　感情と効果音の一致度判定

本節では，同じ説明文とキーワードが付与されている効果音の中から，コンテキスト文とターゲットワードが表す感情に一致する効果音を絞り込む段階について説明する．まず，あらゆるテキストを感情分類するモデルを作成した．次に，コンテキスト文とターゲットワードに対応する感情に対して，効果音が一致するかどうかを判定するモデルを構築した．

4.2.1　感情分類器の構築手法

まず，あらゆるテキストを感情分類するモデルを構築した．以後，このモデルのことを「感情分類器」と呼ぶ．図 4 に感情分類器の詳細を示した．このモデルとして，東北大学の乾研究室が公開する BERT の事前学習モデルを用いる．

BERT をファインチューニングする際，入力となるテキストと正解ラベルのペアデータが必要になる．そこで，梶原らが作成した WRIME データセット [3]を感情分類器に学習させる [10]．WRIME データセットは，文と感情ラベルのペアデータである．ラベルは,「喜び」,「悲しみ」,「期待」,「驚き」,「怒り」,「恐れ」,「嫌悪」,「信頼」の 8 種類である．それぞれのラベルにおいて，評価者 4 人によって，0 から 3 の 4 段階で感情ラベルが付与されたものである．これは，計 80 人の評価者によってラベル付けされた，43,200 件のペアデータセットである．

表 5 に，各感情ラベルにおける，0 から 3 の 4 段階のラベル数を示した．0 のラベルが付与されている感情ラベル数は 276,424 件であり，1 から 3 のラベルが付与されている感情ラベル数の合計は 69,176 件である．よって，全データの中で 0 のラベルが大多数である．

本研究では，この WRIME データセットを用いてモデルをファインチューニングする．ファインチューニングするにあたって，データセットにいくつかの前処理を行った．まず，8 種類すべてのラベルにおいて，評価者 4 人が付けた 0 から 3 の 4 段階の値を平均した．次に，2 以上のラベルをすべて 1，2 より小さいラベルを 0 とした．最後に，8 種類の感情ラベルがすべて 0 となった 20,569 件を削除した．その結果，1 のラベルはすべてで 22,631 件となった．

このデータセットを無作為に 8:1:1 に分割し，それぞれを訓練データ，検証データ，テストデータとした．これらのデータの件数は，訓練データが 18,104 件，検証データが 2,263 件，テストデータが 2,264 件であった．

訓練時には，ファインチューニングを行った [7]．BERT モデルからは入力長の 768 次元のベクトルが出力された．これらのベクトルを Average Pooling したベクトルを FC 層に入力して，8 クラスのマルチラベル分類をした．

4.2.2　一致度判定器の構築手法

次に，コンテキスト文とターゲットワードに対応する感情に対して，効果音が一致するかどうかを判定するモデルを構築した．このモデルを「一致度判定器」と呼ぶ．図 5 に一致度判定器のモデル図を示した．一致度判定器には，Kong らが提案した PANNs [5] を用いる．音声を PANNs に入力すると，2048 次元のベクトルを得ることができる．

一致度判定器を学習させるとき，説明文，キーワード集合，効果音に関する正例と負例が必要となる．

正例　効果音の検索対象データにおける，音の説明文，音のキーワード集合，効果音の 7,547 件

負例　効果音の検索対象データにおける，一致しない音の説明文と音のキーワード集合に対して，効果音をランダムに選択した 7,547 件

上記 2 種類のデータをランダムに混ぜ，無作為に 8:1:1 に分割し，それぞれを訓練データ，検証データ，テストデータとした．これらのデータの件数は，訓練データが 12,075 件，検証データが 1,509 件，テストデータが 1,510 件であった．

訓練時には，ファインチューニングを行った．まず，説明文とキーワード集合をそれぞれ感情分類したベクトルを得る．具体的には，説明文とキーワード集合を感情分類器に入力し，感情ラベル 8 種類に対応した 8 次元ベクトルを 2 つ得る．これ

3：WRIME: https://github.com/ids-cv/wrime

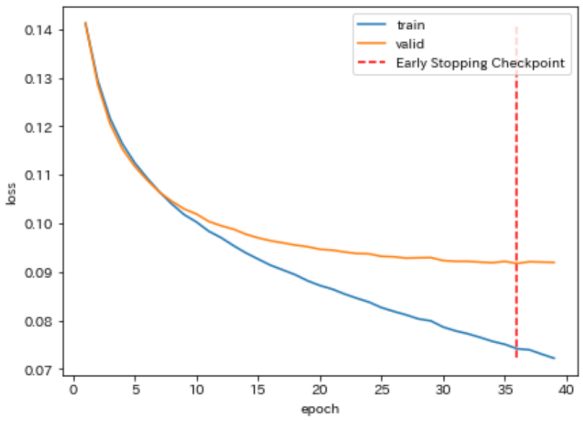

図 6　感情分類器における学習曲線

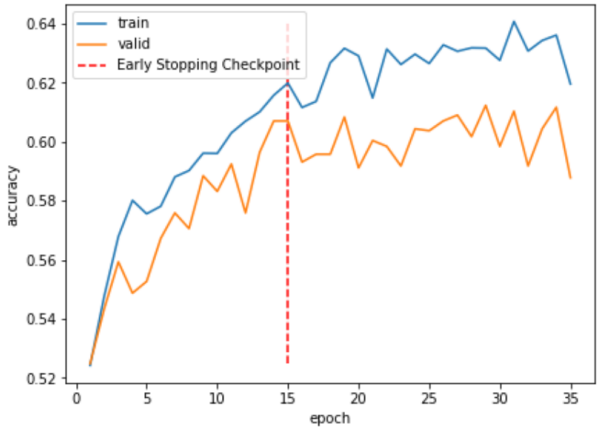

図 7　一致度判定器における正解率の学習曲線

らの 8 次元ベクトル 2 つを結合し，16 次元ベクトルに変換する．次に，前処理として，効果音を 10 秒に収める．10 秒以上の効果音は，10 秒で区切る．10 秒より小さい効果音は，音を繰り返すことで 10 秒に収めた．この効果音を PANNs に入力し，2048 次元ベクトルを得た．これらの 16 次元ベクトルと 2048 次元ベクトルのそれぞれの要素の積をもとめることで，32,768 次元のベクトルに変換した．32,768 次元のベクトルを FC 層に入力し，感情と効果音が一致するか一致しないかを出力した．

4.3　ランキング

4.1 節で得られたテキスト同士の類似度と， 4.2 節で得られた感情と効果音の一致度の積を計算する．この積が大きい順に効果音をランキングした．

5　実　　験

本節では，まず， 5.1 節において，感情分類器における実験と結果について述べる．次に， 5.2 節において，一致度判定器における実験と結果について述べる．

5.1　感情分類器

5.1.1　実　　験

WRIME データセットを感情分類したときのハイパーパラメータは以下の通りである．

- バッチ数：64
- 最適化手法：Adam [4]
- 損失関数：Binary Cross Entropy Loss
- 学習率：1e-5
- epoch：条件により早期終了
- 入力長：256

Early stopping を適用し，36epoch 目に早期終了した．学習時の学習曲線を図 6 に示した．

5.1.2　結　　果

表 6 に，WRIME データセットにおける各ラベルの総数と評価値を示した．「喜び」，「悲しみ」，「期待」，「驚き」，「怒り」，「恐れ」，「嫌悪」，「信頼」といったラベルそれぞれの数は，「期待」ラベルがもっとも多く，次に「喜び」ラベルが多かった．また，「信頼」ラベルに関しても，正解率が 99.47%と高いにもかかわらず，適合率，再現率，F 値すべて 0.00%と低い結果となった．同様に，「怒り」ラベルに関しては正解率が 98.40%ともっとも高いにも関わらず，適合率は 58.33%，再現率は 28.57%，F 値は 38.36%と低い結果となった．これは，他の 6 つのラベルに比べて，「怒り」ラベルと「信頼」ラベルの件数が少ないことが原因であると考えた．

それに対して，「喜び」ラベルに関しては，正解率と適合率は約 70%から 80%となり，再現率と F 値は約 60%となった．これは，「喜び」ラベルの件数がもっとも多いことが原因であると考えた．

感情分類器の推論時には，コンテキスト文とターゲットワードを入力する．

5.2　一致度判定器

5.2.1　実　　験

効果音メタデータと効果音のペアデータを正例と負例の 2 値分類したときのハイパーパラメータは以下の通りである．

- batch size：4
- 最適化手法：Adam [4]
- 損失関数：Cross Entropy Loss
- 学習率：1e-7
- epoch：条件により早期終了

Early Stopping を適用し，15epoch 目に早期終了した．学習時の学習曲線を図 7 に示した．

5.2.2　結　　果

テストデータの正解率は 59.13%であった．

一致度判定器の推論時には，効果音と，コンテキスト文とターゲットワードを感情分類器に入力して出力された 2 つの 8 次元ベクトルを入力する．ファインチューニングされた一致度判定器の出力は，2 次元ベクトルである．この 2 次元ベクトルをソフトマックス層に入れることで，正規化する．

6　出 力 結 果

本節では，(1) から (3) の 3 つの手法による結果を比較した．

表 6　テストデータにおける各ラベルの総数と評価値

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼
件数（件）	5,635	3,584	4,868	2,717	543	2,128	1,849	185
正解率（%）	88.11	88.18	88.96	91.48	98.40	94.18	93.86	99.47
適合率（%）	71.11	52.55	68.74	55.41	58.33	71.84	62.50	0.00
再現率（%）	65.82	38.73	61.54	33.86	28.57	35.40	25.91	0.00
F 値（%）	68.37	44.59	64.93	42.03	38.36	47.44	36.63	0.00

表 7 **「雨」**と**「怒りの雨が降っていた.」**が入力テキストの場合

	(1) テキストの類似度のみに基づく ランキング	(2) 感情と効果音の一致度のみに基づく ランキング	(3) テキストの類似度と, 感情と効果音の一致度の両方に基づく ランキング
Top1	「雨」 「水面に激しく雨が降る音.」	「石」 「大きな石が坂を転がり落ちる音.」	**「雨」** **「水面に激しく雨が降る音.」**
Top2	「雨」 「ガラスに雨が当たる音と, 雨滴がしたたる音.」	「水, 噴水」 「屋外の噴水が水を噴き出す音.」	**「雨」** **「屋根に小雨が当たる音.」**
Top3	「雨」 「セメントに激しく雨が当たる音.」	「雷」 「雷鳴が轟き渡る音.」	**「雨」** **「ガラスやセメントに雨が当たる音.」**
Top4	「雨」 「ガラスや屋根に激しく雨が当たる音.」	「列車, 蒸気」 「列車がベルの音ともに汽笛を鳴らして遠くから近づき, 目の前で停車する音. 列車の横から録った音.」	**「雨」** **「セメントに雨が当たる音と,** **そこへ屋根から雨が流れ落ちる音.」**
Top5	「雨」 「傘に雨が当たる音.」	「水, 雷雨」 「シャーと雨が降る中, 低く雷鳴が轟き渡る音.」	**「水, 雷雨」** **「シャーと雨が降る中,** **低く雷鳴が轟き渡る音.」**

表 8 **「火」**と**「煙草の火が付いた.」**が入力テキストの場合

	(1) テキストの類似度のみに基づく ランキング	(2) 感情と効果音の一致度のみに基づく ランキング	(3) テキストの類似度と, 感情と効果音の一致度の両方に基づく ランキング
Top1	**「火, ガス」** **「ガスがボーッと燃える音.** **やや激しい.」**	「銃」 「44 マグナム銃の撃鉄を起こす音.」	「銃」 「44 マグナム銃の撃鉄を起こす音.」
Top2	**「火, ガス」** **「ガスがボーっと燃える音.」**	「銃」 「コルト式自動拳銃の弾倉を 少しずつ回転させる音.」	「火, ガス」 「ガスがボーッと燃える音. やや激しい.」
Top3	**「火, ガス」** **「ガスにボッと火がつき，燃え上がる音.」**	「銃」 「弾を入れずに コルト式自動拳銃の引き金を引く音.」	「銃」 「引いて 10mm 口径の遊底を開ける音.」
Top4	**「火, ガス」** **「ガスにボッと火がつき，燃え上がる音.」**	「銃」 「引いて 10mm 口径の遊底を開ける音.」	「銃」 「44 マグナム銃の撃鉄を起こす音.」
Top5	**「火, ガス」** **「ガスにボッと火がつき，燃え上がる音.」**	「銃」 「弾を入れずに 拳銃の引き金を引く音.」	「銃」 「弾を入れずに コルト式自動拳銃の引き金を引く音.」

（1） テキストの類似度のみに基づくランキング

（2） 感情と効果音の一致度のみに基づくランキング

（3） 類似度と一致度の両方に基づくランキング

表 7, 表 8 に，効果音のランキング結果の Top 5 件の比較を示した.

表 7 に,「雨」というターゲットワードと「怒りの雨が降っていた.」というコンテキスト文を入力した出力結果を示した.

表 8 に,「火」というターゲットワードと「煙草の火が付いた.」というコンテキスト文を入力した出力結果を示した.

表 7 に示した例では,（1）と（3）の手法において，ランキングされた音のキーワード集合が「雨」のみであった.（2）の手法において，検索結果の第 1 位のキーワード集合は「石」，第

2 位のキーワード集合は「水, 噴水」であった．このような結果となった原因は，コンテキスト文，ターゲットワード，音の説明文，音のキーワード集合それぞれを感情分類器に入力したときの出力ラベルが同一であったためと考えた.「雨」というターゲットワードを感情分類器に入力したときの出力結果は「悲しみ」ラベルであった．同様に,「石」という音のキーワード集合を入力したときの出力結果も「悲しみ」ラベルであった．また,「怒りの雨が降っていた.」というコンテキスト文を感情分類器に入力したときの出力結果は「怒り」と「嫌悪」ラベルであった．同様に,「大きな石が坂を転がり落ちる音.」という音の説明文を入力したときの出力結果も「怒り」と「嫌悪」ラベルであった．

表 8 に示した例では,（1）の手法と（2）の手法でランキングされた音のキーワード集合は「火, ガス」と「銃」に二極化した．これは，単純にテキスト同士の類似度を計算する手法と，テキストの感情と効果音の類似度を計算する手法での差があるためと考えた.（1）の手法では，ターゲットワードである「火」と同一の語が含まれているキーワード集合と類似度があるとみなされた.（3）の手法では，コンテキスト文，ターゲットワード，音の説明文，音のキーワード集合それぞれを感情分類器に入力したとき，出力された感情ラベルはすべて異なっていた．コンテキスト文である「煙草の火が付いた.」を感情分類器に入力したときの感情ラベルは「驚き」ラベルであり,「44 マグナム銃の撃鉄を起こす音.」という音の説明文を入力したときの感情ラベルは「悲しみ」ラベルであった．また，ターゲットワードである「火」を感情分類器に入力したときの感情ラベルは「怒り」ラベルであり,「銃」という音のターゲットワードを入力したときの感情ラベルは「喜び」ラベルであった．よって，効果音を PANNs に入力し，出力されたベクトルによって,（1）とは異なるランキングの結果となったと考えた．

7 まとめと今後の課題

本研究では，同じ説明文やキーワード集合が付与されている効果音の中から，コンテキスト文やターゲットワードと感情が一致する効果音を絞り込んで検索する問題に取り組んだ．効果音検索の入力は，たとえば「怒りの雨が降っていた.」のようなコンテキスト文と,「雨」のようなターゲットワードである．それに対して，雨が激しく降りしきる効果音が出力される．

本研究の課題は 2 つである．まず，同じ説明文とキーワードが付与されている効果音を検索する課題である．次に，コンテキスト文とターゲットワードが表す感情に一致する効果音を絞り込んで検索する課題である．出力結果を 6 節に示した．

（1） テキストの類似度に基づくランキング

（2） 感情と効果音の一致度に基づくランキング

（3） 類似度と一致度の両方に基づくランキング

今後は,（1）から（3）の 3 種類の手法を比較することにより，出力結果を評価する予定である．

謝　　辞

本研究は JSPS 科研費 JP21H03775，JP21H03774 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Alexei Baevski, Henry Zhou, and Michael Mohamed, Abdelrahman Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 2020 Conference on Neural Information Processing Systems*, pp. 1–19, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4920–4928, 2019.

[3] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *Proceedings of the 2018 International Conference on Machine Learning*, pp. 2410–2419, 2018.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference on Learning Representations*, 2015.

[5] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. *PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.

[6] Mariana J. Lopez and Sandra Pauletto. The sound machine: a study in storytelling through sound design. In *Proceedings of the 2010 Audio Mostly Conference: A Conference on Interaction with Sound*, pp. 1–8, 2010.

[7] Alice Lucas. *Deep perceptual losses and self-supervised fine-tuning for image and video super-resolution*. Northwestern University, 2020.

[8] Yi Luo and Nima Mesgarani. Conv-tasnet: surpassing ideal time-frequency magnitude masking for speech separation. In *Proceedings of the 2019 IEEE/ACM Transactions on Audio, Speech and Language*, pp. 1–12, 2019.

[9] Sophia C. Steinhaeusser, Philipp Schaper, and Birgit Lugrin. Comparing a robotic storyteller versus audio book with integration of sound effects and background music. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 328–333, 2021.

[10] Kajiwara Tomoyuki, Chu Chenhui, Takemura Noriko, Nakashima Yuta, and Nagahara Hajime. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 2095–2104, 2021.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 2017 International Conference on Neural Information Processing Systems*, pp. 5998–6008, 2017.

[12] 山田侑樹, 櫨山淳雄, 小川雄太郎. OSS プロジェクトの Issue 議論内容に対する BERT および AutoML を用いた文章分類の提案. 人工知能学会第 34 回全国大会論文集, pp. 1–4, 2020.

[13] 内藤勝太, 白松俊. Web 議論における BERT を用いた関連情報推薦エージェント. 情報処理学会第 82 回全国大会講演論文集, pp. 637–638, 2020.