Vol.2022-IS-159 No.8
2022/3/7

自然対話音声に含まれる抽象的感情の AI による推定

陳 春陽[1]　近藤 公久[1]

概要：自然かつインテリジェントに人間とコミュニケーション可能なロボットやエージェントの実現には対話相手の感情状態の推定が必須であるが技術的には課題が多く存在する．近年の AI 技術の発展はこの課題解決のためにも有効であると期待されており，すでに「怒り」や「喜び」などの基本的な感情を含む音声かどうかを識別する技術が開発され，一部は実用に供している．しかし，コミュニケーションにおける相手の様々な感情の推移を推定することは容易ではない．そこで本研究では，宇都宮大学パラ言語情報研究向け音声対話データベースの発話に対する 6 つの感情状態の評定値を用いて，抽象的感情を推定する AI モデルの検討を行った．その結果，AI が一定以下の誤差範囲で感情を推定可能であることが示された．

キーワード：ディープニューラルネットワーク，感情推定，自然言語処理

Voice emotion estimation AI model of abstract emotions contained in natural dialogue

CHEN CHUNYANG[†1]　TADAHISA KONDO[†1]

Keywords: Deep Neural Network, emotion estimation, Natural language processing

1. 背景

5G と IoT（Internet of Things）の普及とテレワークの増加などの背景の中，人々は機械と向き合う時間が増え，人とのコミュニケーションの減少は社会的問題になっている．人々は対人コミュニケーションを通して自己概念や自己評価を形成する[1]．自己概念については，受け手から望まれる自己像を示すだけでなく，自分自身の考え等を同時に示すことが精神的な健康と正の相関を示すことが報告されており，健全な人間として欠かせない概念だと考えられる．そのため，人とスムーズにコミュニケーションできる AI は様々な分野においてニーズが高まっている．

近年，機械と音声によるコミュニケーションをベースとした様々な音声インタラクションサービスが実用になっている．音声アシスタントではアップル社の Siri，マイクロソフト社の Cortana や Google の Google アシスタントなどがあり，スマートスピーカーではアマゾン社の Echo，Google の Google Home，アップル社の Homepod などがある．このように，コンピューターとスマートフォンに搭載されている音声アシスタントから，独立した音声インタラクション製品まで，世界的な大手 IT 企業が激しく競争していることからも，音声インタラクション技術の重要性や期待がうかがえる．

1.1 音声アシスタントの利点と問題点

音声アシスタントの利便性の最も重要なポイントは，従来の GUI では実現不可能な速さと操作の容易さにあるといえる．たとえば，目覚まし時計を設定するにはボタンなどを何度も押す必要があるが，音声アシスタントでは「明日○○時に起こして」と声をかけるだけでよく，手指による操作が不要なため他の仕事をしながら命令することができる．ジェスチャーと声でコミュニケーションしていた時代の後に文字の発明があり，赤ちゃんは最初に話すことを覚えてから文字を学ぶなど，音声コミュニケーションが人間にとって基本的なものと考えることは自然である．

一方で，現存する AI アシスタントでは音声認識誤りも多く，ユーザーが一単語ずつはっきりと喋らなければ理解できないこともよく経験する．また発話内容の曖昧性から，間違った意味にとらえられ，うまく指示ができない場合もある．さらには自由対話ができる AI は少なく，会話が弾まず，ユーザーが会話を諦めるケースもよくある．

自由対話の難しさは，相手の発話の意図の理解にある．同じ言葉でも，話す環境や発話者の感情によって言葉の意味が変わり，それに対するレスポンスを変える必要がある．例えば，同じ「hands up!」でも，ライブなどテンションが高いシチュエーションにおける場合と，路上で脅迫的な声で言われた場合では全く受け取る意味が異なる．この例はかなり特殊な状況の違いではあるが，対話における微妙な相手の感情の推移や発話の意図やニュアンスの理解は，スムーズなインタラクションには欠かせない．音声には言語情報以外の情報が含まれており，そこから対話相手の感情を認識し，それに相応しい反応が選べる AI アシスタントへの期待はますます高まる．

1.2 AI による感情推定

Picard[2]は，「自然かつインテリジェントに人間とインタラクトできるコンピューターには，少なくとも感情を認識して表現する能力が必要である」と述べている． Picard が

感情推定 AI（Affective Computing）を初めて提唱して以来．多くの研究者らが，表情，言語，心拍，呼吸などから人間の感情を解析する感情推定 AI の開発を進めてきた．その中で，「怒り」「喜び」などの基本的な感情の推定には 80%以上の精度が確認できている研究も少なくはない．それでも，「商用化はまだ遠い」，「信用できない」などの指摘もある．その原因は，人々が現実の生活にうまく言えない曖昧な気持ちを感じることが多くあり，むしろ簡単に基本感情に分類できることは稀なことであるからといえる．Ortony ら[3][4]は，「感情とは，人が心的過程のなかで行う様々な情報処理のうちで，人，物，出来事，環境について行う評価的な反応である」とし，対象を「良し一悪し」「危険一安全」「有用一有害」「好き一嫌い」などの軸に基づく評価を行い，その評価に基づいたあらゆる感情的な反応を「表現」と述べた．

Russell が提唱した感情円環モデル[5]では，感情を「快楽度」と「覚醒度」の二軸で分解して表現する． 基本感情を「快楽度」「覚醒度」などの「分子」と例えると，その中に「良し一悪し」「危険一安全」「有用一有害」「好き一嫌い」などの「原子」が存在していると考えることができ，もし「分子」について感情推定できれば，人間の自然対話の中の細かい微妙な感情も，その「分子」の組み合わせによって，もっと細かく表現できると考えられる．本研究ではこういった感情「分子」を「抽象的感情」と呼ぶ．

1.3 目的

人間のように会話するコンピュータやロボットは SF のような夢の世界のことではすでになくなってきており，音声を聞き取る音声認識技術や人らしい音声で喋る音声合成技術，および，翻訳技術は近年の AI 技術により大きな進歩をとげている．しかし，「人間の気持ちを読み取る」など，スムースなコミュニケーションの実現については上述のとおり未だ発展途上である．

人間の自然対話の中の微妙な感情を精確に読み取るには基本感情だけではなく，抽象的感情に着目する必要がある．しかし，対話の中で時々刻々と変化する微妙な感情の変化をコンピュータによって捉えるのは未だ困難であり，先行研究も少ないのが現状である．本研究では基本感情を分類・認識する AI モデルを発展させ，ディープニューラルネットワーク（DNN）のクラス分類と線形回帰分析の二つアルゴリズムでモデルを構築し，感情の「分子」と考えられる抽象的感情を推定可能にする AI 手法について検討を行った．

2. 音声データセット

本研究では，宇都宮大学パラ言語情報研究向け音声対話データベース（UUDB：Utsunomiya University Database）[6]を用いた．UUDB は，自然で表情豊かな音声対話を集めたコーパスであり，最大の特徴の 1 つは，全ての発話に対し，話者がどのような感情状態であったと感じられるかを主観的に評価したラベルが付与されているところにある．このラベルは，「快楽度」「覚醒度」「支配度」「信頼度」「関心度」「肯定度」の 6 項目であり，各発話毎に付与されている，評定値は 3 人によって評価されたものである．

パラ言語情報のラベリングの安定性は既によく調べられており[7], UUDB のパラ言語情報ラベルは一定の範囲で信頼できる．UUDB では 6 項目のラベルを感情の抽象次元と呼んでいるが，本研究はそれらを感情の「分子」として考え，これを抽象的感情とみなした．

UUDB には二人の対話が 27 対話（セッションと呼ぶ）が収録されており，全発話数は全体で 4840 文，各セッションの平均発話数は約 180 文である．図 1 は UUDB の各感情の評定値分布であり，横軸は感情の評定値，縦軸はデータ（発話）の出現頻度である．「支配度」以外の各感情では，評定値が 2 未満と評定値が 6 以上のデータが極めて少なく，概ね正規分布に近いといえる．一方，「支配度」では，評定値が 3 から 6 のデータが平均的にあり，2.5 あたりにピークがある．また，全体的には山が 2 つ存在するような形にもみえる．

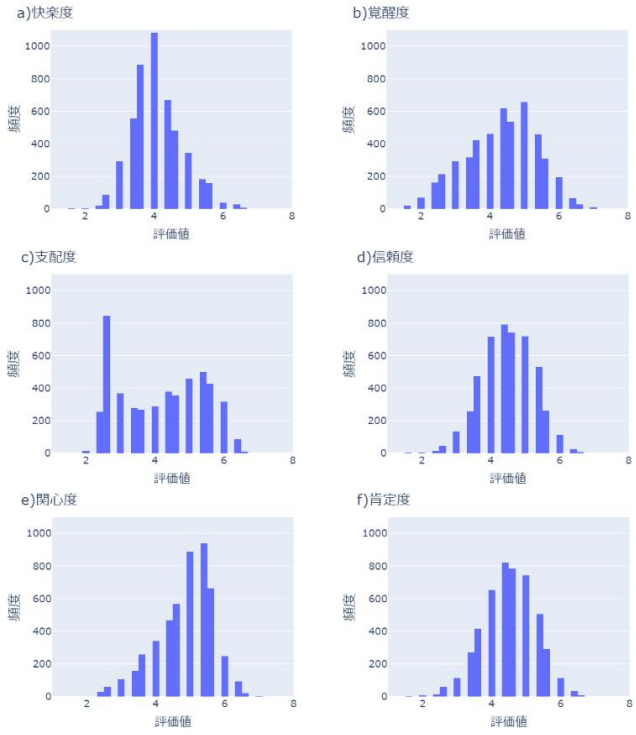

図 1 評定値のヒストグラム

3. 実験 1

本実験では，Shah ら[8]の学習モデルを参照し，UUDB**[6]**を用いて，自然対話における 6 種の抽象的感情について二値分類を行う．Shah らは「楽しみ」「怒り」「落ち着き」「悲しみ」「驚き」「中立」「嫌悪」「恐怖」の 8 種の基本感情について感情推定(分類)を行い，約 70%の精度を得た，その結果と本実験の結果を比較し，同じ学習モデルが抽象的感

情を対象とした感情推定(分類)に適用できるかどうかを確認する．

3.1 実験手法

本実験は図 2 のように「データの整理」「音響特徴値の抽出」「教師フラグの付与」「テストデータの準備」「モデルのトレーニングとテスト」の 5 つの手順で行った．

(1) データの整理

UUDB は評定値 1~7 のデータであり，二値分類を行うために，1 と 0 の二つのグループに分ける作業を行った．二値に分ける基準を決めるために，UUDB の全発話に対する評定値の平均値と中央値を求めた(表 1). 前章で示したとおり，「支配度」以外の「快楽度」「覚醒度」「信頼度」「関心度」「肯定度」の評定値分布は正規分布に近く，表 1 においても平均値と中央値が近いことが確認できるため，平均値を基準に二値(1 :曖昧感情の度合いが低いグループと 0: 曖昧感情の度合い低いグループ）に分類し，「支配度」は評定値の中央値を基準に分類した．

表 1　パラ言語情報評定値の平均値と中央値

	平均値	中央値
快楽度	**4.09**	4.00
覚醒度	**4.33**	4.33
支配度	4.07	**4.33**
信頼度	**4.36**	4.33
関心度	**4.95**	5.00
肯定度	**4.61**	4.67

(2) 音響特徴値の抽出

グループ 1 とグループ 0 にあるすべての音声データから MFCC，Mel spectrogram，spectral contrast，chromagram，Tonnetz representation の 5 種類の音響特徴量を抽出した．この 5 種類の音響特徴量は Issa ら[9]が提案したものである．以来，Shah らの研究[8]をはじめ，数多くの研究に使用されて検証が行われてきたため，本研究でも同じ特徴量を用いて実験を行った．音響特徴量の抽出には，各発話データを窓サイズ 2048（おおよそ 10ms），スライド幅 512（1/4 窓サイズ）のハニング窓をかけたものに対して LibROSA を用いて得られた値を平均したものであり，この値をその発話の特徴値とした．それぞれの特徴量について以下にまとめる．

・MFCC（メル周波数ケプストラム係数）

MFCC は Stevens ら[10]によって 1937 年に提案された音高の知覚的尺度であり，人間の聴覚に基づき，メル尺度の差が同じであれば，人間が感じる音高の差が同じになることを意図して，周波数を式（1）によってメル尺度に変換したものである．

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \qquad (1)$$

MFCC は計算が簡単で識別能力が高いなどの特徴があるため，音響の幅広い分野で使用されており，音響認識システムを開発する際に音響特徴量の第一候補とされている[11][12]. 感情推定の分野でも，MFCC は「怒り」や「喜び」などの基本感情の間で大きな差異が生じることが発見されている[13].

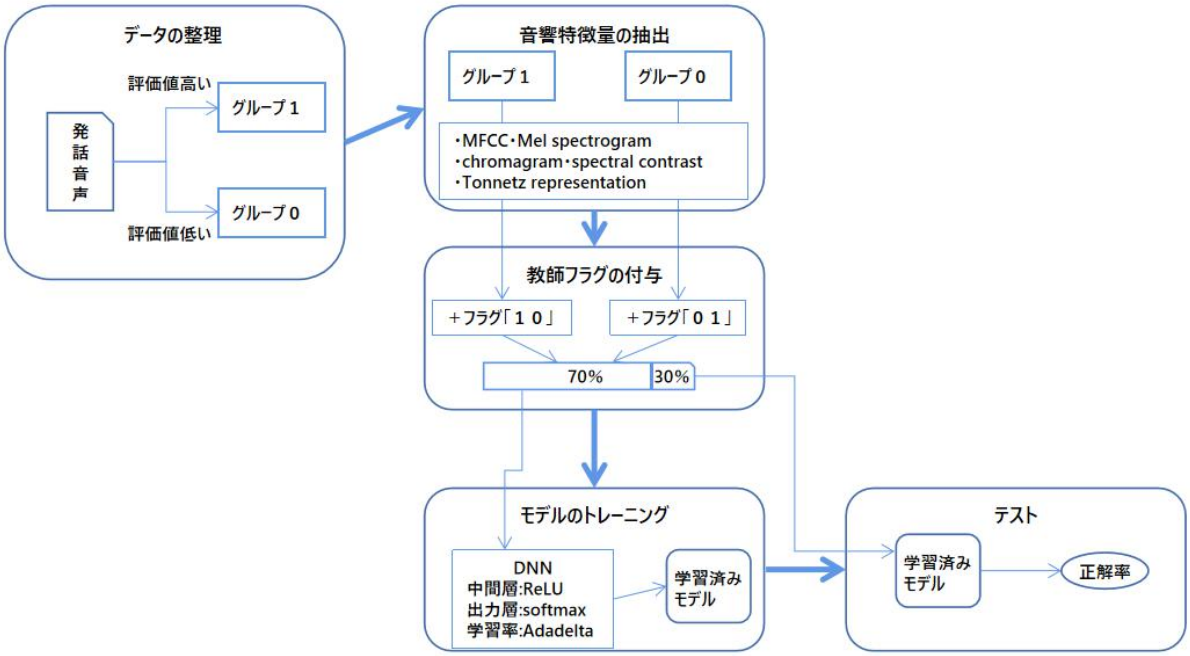

図 2　実験 1 フロー図

・Mel spectrogram（メルスペクトログラム）

Mel spectrogramはMFCCと同じ原理で計算可能なものである．Mel spectrogramから対数を取り，離散コサイン変換をすればMFCCとなる．離散コサイン変換は非可逆圧縮のためデータの損失が発生する．ディープランニングはなるべく多くのデータを要するため，音声感情推定にはMel spectrogramを音響特徴量として使用するのが近年主流となっているが，Venkataramananの研究によると，両方同時に使用した感情推定の精度は片方だけ使用した場合より高い精度が確認できているため[14]，本研究でも両方とも使用する．

・spectral contrast（スペクトルコントラスト）

spectral contrastはJiangら[15]によって2001年に提案され，音響データの各周波数領域の相対的スペクトル特性を表している音響特徴量である．spectral contrastを使用した音楽のジャンル分類学習はMFCCを使用する場合より高い精度が確認されている．

・chromagram（クロマグラム）

chromagramは音色や楽器を頻繁に切り替えている音楽に対しても，精確に高調波特徴と旋律特徴をキャプチャすることができるため，音楽ジャンルや楽器の分類学習によく使われている音響特徴量の一つである[16][17].

・Tonnetz representation（Tonnetz 表現）

Tonnetz表現はLeonhard Eulerによって1739年に最初に提案された，音のピッチ関係を平面に表現する音響特徴量である．現在はHarteらの改良によって，和音の変化と高調波特徴の変化も表現できる音響特徴量である[18].

MFCCとMel spectrogramは音声感情分類に普遍的に使用されているが，spectral contrast，chromagramとTonnetz表現は主に音楽のジャンル分類や音楽感情の研究に使われている音響特徴量である，Issaら[9]では5種類の音響特徴量で学習精度71.61%の感情推定ができた，Shahら[8]の学習モデルでも同じ音響特徴量で70%を越えた学習精度が確認できた．Venkataramanan[14]はMFCCとMel spectrogramだけでの感情推定での66%の精度と比べて，spectral contrast，chromagramとTonnetz representationの追加は確かに学習精度を向上できることを示した．以上はすべて英語コーパスRAVEDESS[19]を対象とした感情推定実験である．

(3) 教師フラグの付与

手順2で抽出した音響特徴量にone-hot表現で教師フラグを付与する，音響特徴量を抽出した音声データが所属するグループのグループ番号により，グループ1の場合はフラグ「0 1」，グループ0の場合はフラグ「1 0」を付与する．one-hot表現（one-hotベクトル）とはベクトルの全ての要素のうち一つだけ1で他は0になっているベクトルを意味する．クラス分類や自然言語処理に使われるデータの前処理手法である．すべてのクラスを平等に扱えるメリットがあるが，クラスの数が多すぎると，ベクトルの次元が増え，分析手法によっては，計算に非常に時間がかかったり，そもそもまともな分析ができない場合もある．

(4) モデル学習（トレーニング）

UUDBの発話音声の全体27セッションの内，20セッションをトレーニングデータとして使用し，発話数はUUDB全体の4840発話中の3407発話，おおよそ70%であった．

学習モデルは，Shahら[8]のモデルを参照した．実装においては，モデル構築と学習にはKERASを用いた．学習モデルは中間層が3層で構成されたDNNを使用した．各中間層のノード数を200，400，200とし，活性化関数はランプ関数（ReLU）を使用した，各層間のドロップアウトは0.2とした．出力層では，入力した値を0~1の確率的な値に変換するソフトマックス関数（softmax）を活性化関数とし，最適化アルゴリズムは勾配更新の固定移動ウィンドウに応じて，学習率を調整するAdadelta関数を使用した．トレーニングデータを以上の学習モデルに学習させ，学習済みモデルを構築した．

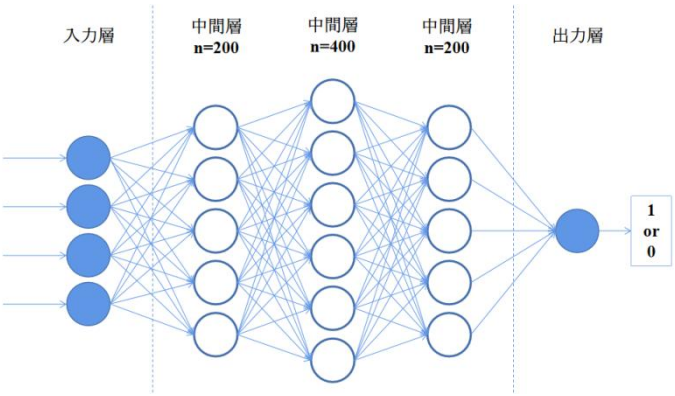

図 3　学習モデル図

(5) テスト

手順(4)のモデル学習に用いられなかったUUDBの1433発話，全体のおおよそ30%をテストデータとして使用した．テストデータの分布はデータ全体(図1)とほぼ同じであることを確認した．

テストデータを学習済みモデルによって各抽象的感情を含むか否かについて二値分類を行った，出力値は「10」もしくは「01」の二値である．「10」の場合は「快」「覚醒」「支配」「信頼」「関心」「肯定的」の度合いが高いグループに分類したことを意味し，「01」の場合は「不快」「睡眠」「服従」「不信」「無関心」「否定的」の評定値の低い感情に分類したを意味する．

3.2 実験1結果と考察

本実験はShahら[8]の実験手法を参照し，同じ音声特徴

値と同じ学習モデルで日本語コーパスを使い，より複雑な抽象的感情に対して二値分類を行った．そのテスト結果を表 2 と図 4 で示した．

表 2 は各抽象的感情の各評価区間と全体の正解率を数字で示したものであり，区間は評価値 1 間隔で取ったものである．「覚醒度」と「関心度」の評定値の高い（5~6）の部分と「支配度」の評定値の低い（2~3）では 80%を越える高精度が得られたが，「覚醒度」の評定値の 2~3 や「関心度」の評定値 3~4 では正解率が 30%前後しか得られなかった区間がいくつかある，一番正解率の低い区間は「覚醒度」の評定値 3~4 の 11.1%である．

表 2　実験 1 分類結果

	正解率(データ総数)					
評価値	2<=x<3	3<=x<4	4<=x<5	5<=x<6	6<=x<7	全体
快楽度	47.1%(17)	36.9%(436)	75.6%(628)	65.0%(317)	52.9%(34)	60.6%(1433)
覚醒度	31.9%(116)	11.1%(306)	66.2%(420)	93.8%(452)	83.6%(134)	61.8%(1433)
支配度	86.0%(414)	74.9%(271)	32.7%(202)	33.1%(341)	42.4%(205)	57.6%(1433)
信頼度	28.6%(7)	40.1%(167)	69%(621)	60.6%(556)	55.6%(81)	61.5%(1433)
関心度	62.2%(37)	31.8%(176)	50.5%(384)	77.6%(662)	81.3%(172)	64.8%(1433)
肯定度	80.0%(5)	57.9%(152)	48.8%(633)	61.0%(575	54.5%(68)	55.1%(1433)

図 4 は正解，不正解と全体の割合を表す折れ線グラフであり，青い線は正解，赤い点線は不正解，緑破線は全体のデータ数を表している．「快楽度」と「信頼度」は中央の正解率が一番高く，両極に徐々に落ち行き，評価値 3~4 の間の正解率が一番低い．「覚醒度」と「関心度」は同じく評価値が高いところで精度がよく，評価値の低いところで精度が低い．「支配度」のデータ全体の分布は他の感情と異なっているため，データ数が多い評価値 2~3 の正解率が一番高くなったものと考えられる．最後に「肯定度」では各区間で平均的な正解率が確認されている．

各感情の全体的な正解率は 60%前後となった．先行実験の 75%前後の正解率と，Shah ら[8]の八次元分類の 70%の正解率と比べると，チャンスレベル 50%の二値分類にとって決して良いパフォーマンスとは言えない．

原因としては用いた抽象的感情は，その感情を含むか否かの強制的な二値化にあると考えらる．抽象的感情の強いから弱いの境界線は不明瞭で曖昧であり，各感情の変化は連続的である．このため，平均値や中央値によって分けた分類においては学習精度に大きく影響を与えた可能性がある．また，曖昧であると考えられる基準値前後のデータを削り，両極のデータだけを用いた場合でもそれほど大きな改良は見られなかった．そこで，線形で連続的な感情の変化の基づいて，線形回帰の学習手法で，精確の高い感情推定ができるではないかと考え，実験 2 を実施した．

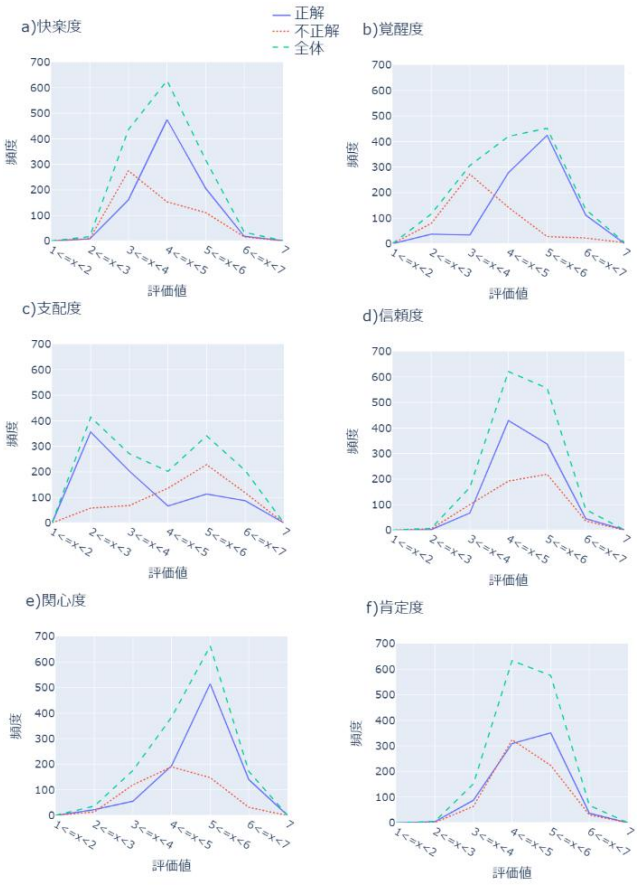

図 4　実験 1 分類結果

4. 実験 2

4.1 実験目的

前章では，抽象的感情を DNN のクラス分類の手法で，感情を含むか否かについて二値分類を行った，芳しい精度を得られなかった原因は感情の変化は連続的であり，高いと低いの境界線が不明確であるためと考えた．

「連続的」という特徴から線形回帰を連想し，本実験は UUDB の発話音声を用いて，線形回帰の学習手法により，発話音声に含まれる各抽象的感情の連続値推定を行い，その結果を評価する．

4.2 実験手法

本実験は実験 1 と同じく「データの整理」「音響特徴値の抽出」「教師フラグの付与」「モデルのトレーニング」「テスト」の 5 つの手順で行った，修正したポイントは主に「データの整理」と「モデルのトレーニング」と「テスト」にある(図 5).

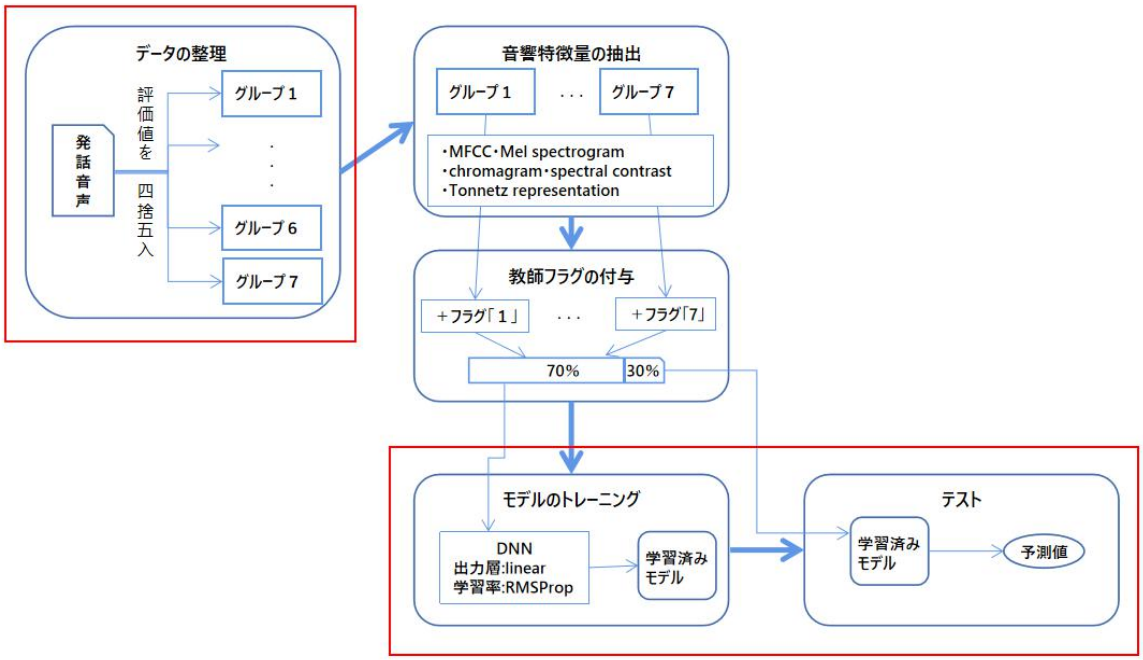

図 5　実験 2 フロー図

(1) データの整理

本実験は連続値推定を行うため，音声データを各曖昧感情の平均評定値の小数第 1 位で四捨五入した整数値により，1 から 7 の 7 グループに分けた．

(2) 教師フラグの付与

音響特徴量を抽出した音声データが所属するグループのグループ番号により，1 から 7 の教師フラグを付与した．

(3) 音響特徴量の抽出

実験 1 と同じ MFCC，Mel spectrogram，spectral contrast，chromagram，Tonnetz representation の 5 種類の音響特徴量を抽出した．

(4) モデルのトレーニングとテスト

実験 1 と同じ UUDB の発話音声の全体 27 セッションの内，20 セッション（全体の約 70%）をトレーニングデータとして使用した．

実験 1 と同じく，モデル構築と学習には KERAS を用い，学習モデルは中間層が 3 層（各中間層のノード数を 200，400，200）で構成された DNN を使用した．入力刺激値をそのまま出力刺激値とする線形関数（linear）を活性化関数とし，最適化アルゴリズムは勾配の大きさに応じて学習率を調整する RMSProp 関数を使用した．

(5) テスト

実験 1 と同様に，手順（4）のモデル学習に使用しなかった UUDB の発話音声（約 30%）をテストデータとして使用した．

テストデータを学習済みモデルによって各抽象的感情の評定値の連続値推定を行った．

4.3 実験 2 結果と考察

得られた学習モデルでテストデータの抽象的感情の評定値の推定を行った結果を表 3 と図 6 に示す．

表 3 には，各抽象的感情の推定結果の誤差として「平均二乗偏差（RMSE）」と「平均絶対誤差（MAE）」，モデルを評価するための「平均二乗偏差と平均絶対誤差の比」と相対誤差を示している．また，図 6 には，各抽象的感情に対し，横軸が UUDB の評定値，縦軸を推定値とした散布図と Lowess 平滑化したトレンドラインを示し，各散布図の上に各評定値の区分毎のデータ数のヒストグラムを示している．

表 3　実験 2 推定結果誤差

	推定結果誤差			
	RMSE	MAE	RMSE/MAE	相対誤差
快楽度	0.61	0.49	1.25	11.9%
覚醒度	1.05	0.88	1.19	21.9%
支配度	1.11	0.89	1.25	23.3%
信頼度	0.86	0.71	1.20	15.4%
関心度	0.76	0.60	1.26	13.9%
肯定度	0.74	0.59	1.25	12.7%

RMSE:平均二乗偏差
MAE:平均絶対誤差

表 3 から，「快楽度」「信頼度」「関心度」と「肯定度」では，平均二乗偏差（RMSE）が 0.6~0.8 程度で全体的に良好な精度と考えられる結果が確認され，「覚醒度」と「支配度」では，RMSE が 1.05 と若干精度が下がる結果といえる．また図 6 から，トレンドラインはすべて斜め右上となっているため，学習モデルは各感情について大まかな推定が出来，すべての抽象的感情の各評定値の推定について大きな誤差が発生することは少ないといえ，「不快」や「睡

眠」などのネガティブな感情を「快」や「覚醒」などのポジティブな感情に推定するような大きな誤差が発生することもほとんどみられない．また，RMSE と MAE（平均絶対誤差）の比が$\sqrt{(\pi/2)} \cong 1.25$ に近い値であるとき，モデルはデータの大まかな特徴を表現できているとされる[20]．この指標に従えば，本結果のすべての抽象的感情における RMSE と MAE の比は 1.20~1.25 の範囲であり，良好なモデルが構築できたと考えられる．

感情毎にみると，誤差が一番小さく，散布図の分散幅も小さく，トレンドラインから最も対角線に近い「快楽度」が，すべての抽象的感情の中で一番精度が高い結果を得ていると考えられる．「肯定度」と「覚醒度」のトレンドラインも似たような傾きに見えるが，「覚醒度」の RMSE は他より大きく，必ずしも高い推定精度が得られたとはいえない．「信頼度」と「関心度」では，表 3 の誤差の結果からは全体的に良好な精度が確認できたが，図 6 から，評定値 5 以上では比較的高いパーフォーマンスが得られているものの，4 以下ではやや低いパフォーマンスとなっているようにみえる．「支配度」の学習精度が一番芳しくないと考えられ，全体のデータが評価値の低い方に偏っているのが原因ではないかと考えられる．

5. 結論

本研究はディープニューラルネットワーク（DNN）のクラス分類と線形回帰分析の二つアルゴリズムでモデルを構築し，感情の「分子」と考えられる「快楽度」「覚醒度」「支配度」「信頼度」「関心度」「肯定度」の 6 種類の抽象的感情を推定する AI 手法の検証を行った．

実験 1 では Shah ら[8]が構築した，「怒り」「喜び」などの基本感情分類で高い精度を確認できたモデルを参照し，発話音声が抽象的感情を含むか含まないかの二値分類学習を実施したが，Shah らの 8 次元から 2 次元に簡略化した同じモデルであっても高い精度を得られなかった．その原因は抽象的感情は基本感情と比べ，感情の強いと弱いの境界値が微妙で不明確であるためと考えられる．つまり，UUDB の評定値は連続的であり，平均値や中央値で分けた場合には，中央付近のデータは類似した音声特徴量を持っているため，学習精度に大きく影響を与えたのではないかと考えられる．

実験 2 では，UUDB の連続的な評定値の特徴を感情の強いから弱いの連続変化として学習し，線形回帰分析の方法でモデルを構築して検証を行った．すべての 6 種類の感情について一定以上の精度で推定可能であり，「不快」や「睡眠」などの度合いの低い値を真逆の「快」や「覚醒」などの度合いの高い値に推定するような大きな誤差はほとんどないことが確認できた．特に「快楽度」と「肯定度」は他の感情より，比較的良好な推定結果が確認できた．しかし，本モデルによる推定結果では，正解値（対角線）から離れたデータも少なくはない．学習データの量を増やすことや，より多くの感情を表す音響特徴量を増やすことなど，さらなる精度向上のための方法が考えられる．また，発話の前後関係や対話者間の影響などについてもモデルに入れることも，より精度の高い感情推定 AI モデル構築が期待できる．

一方で，UUDB の評定値は評定者によって評定されたものであり，評定者は言葉の意味，前後の発話の関係と対話者間の影響を含めて考慮し，評定値を付けたものであるのに対し，本学習モデルでは音声特徴量のみからの推定である．そこで，言語情報も含めたモデル化も今後の検討としたい．ただし，抽象的感情は，言語表現としてあからさまな表現として表出されることはほとんどなく，感情語をキーとした推定が容易にできるものではない．対話者間の関係の推定も容易ではないため，課題は多い．さらに，先行する対話の状態遷移から次の発話，あるいは対話の先の感情状態の予測も視野にいれて，改良を続ける予定である．

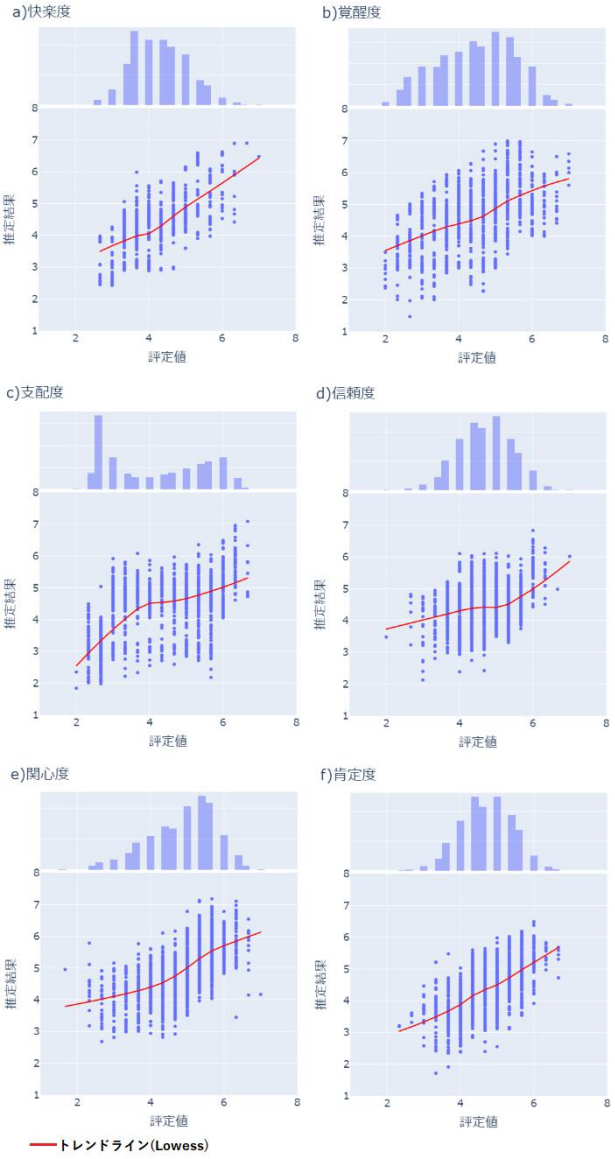

図 6　実験 2 結果散布図

参考文献

[1] 小野 美和: “対人コミュニケーション研究における課題についての一考察”, 愛知淑徳大学論集-福祉貢献学部篇-第 7 部,2017
[2] Picard, R.W.(2000): “Affective Computing”
[3] Ortony, A., Clore, G.L., Collins, A.(2000): “The Cognitive Structure of Emotions”
[4] 大平英樹 (2010) : 感情心理学・入門
[5] Russell, J. A. (1980): “A circumplex model of affect. Journal of Personality and Social Psychology”, 39(6), 1161–1178.
[6] 森 大毅 (2008): “宇都宮大学パラ言語情報研究向け音声対話データベース(UUDB)”, 国立情報学研究所 音声資源コンソーシアム. (データセット). https://doi.org/10.32130/src.UUDB
[7] 森 大毅, 相澤 宏, 粕谷 英樹: “対話音声のパラ言語情報ラベリングの安定性” 日本音響学会誌, 61(12), pp.690-697, 2005
[8] Shah, A., Firodiya, S.: Audio Sentiment Analysis after a Single-Channel Multiple Source Separation,Indiana University Bloomington(2020)
[9] Issa, D.; Demirci, M.F.; Yazici, A.: “Speech emotion recognition with deep convolutional neural networks”. Biomed. Signal Process. Control 2020, 59, 101894.
[10] Stevens, S.S., Volkmann, J., Newman, E.B.: “A scale for the measurement of the psychological magnitude pitch”, The Journal of the Acoustical Society of America 8, 185 (1937)
[11] Davis, S., Mermelstein, P.: “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. IEEE Transactions on Speech Acoustic Processing, 1980 ,28, pp. 357~366
[12] Skowronski, M.K., Harris, J.G.: “Increased MFCC filter bandwidth for noise-robust phoneme recognition”. IEEE International Conference on Acoustics, Speech, and Signal Processing.2002, 1, pp.801~804
[13] Li, H., Xu, X.L., Wu, G.X., Ding, C.Y., Zhao, X,M,: “Research on speech emotion feature extraction based on MFCC”. Journal of Electronic Measurement and Instrumentation, 2017, 31-3,pp. 448~453
[14] VENKATARAMANAN, K.: “RAJAMOHAN, Haresh Rengaraj. Emotion Recognition from Speech”. CoRR.2019
[15] Jiang, D.N., Lu, L., Zhang, H.J., Tao, J.H., Cai, L.H.: “Music type classification by spectral contrast feature”, In Multimedia and Expo, 2002. ICME’02. Proceedings. 2002 IEEE International Conference on, vol. 1, pp. 113-116. IEEE, 2002.
[16] Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. C., Richardson,P., Scott, J., et al.: “Music emotion recognition: A state of the art review”, Proc. 11th Int. Soc. Music Inf. Retrieval Conf. (ISMIR), pp. 255-266, 2010.
[17] Ellis, D. P. W., Poliner, G. E.: “Identifying ‘cover songs’ with chroma features and dynamic programming beat tracking”, in Proc. Int. Conf. Acoustic, Speech and Signal Processing, Honolulu, HI, 2007.
[18] Harte, C., Sandler, M., Gasser, M. (2006): “Detecting Harmonic Change in Musical Audio.” In Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia (pp. 21-26).
[19] Livingstone, S.R., Russo, F.A. (2018): “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English”. PLoS ONE 13(5): e0196391.
[20] Chai, T., & Draxler, R. R.: “Root mean square error (RMSE) or mean absolute error (MAE)- Arguments against avoiding RMSE in the literature”, Geosci. Model Dev., 7, 1247–1250, (2014).