［研究助成（C）］

マルチメディアデータを用いたディープラーニングによる要介護者の感情認識

Emotion recognition using deep learning model with multi-media data

2197013

研究代表者　大阪大学大学院　基礎工学研究科　特任助教　羅　兆傑
（助成受領時：神戸大学　システム情報学研究科　博士課程）

共同研究者　神戸大学　システム情報学研究科　教　授　滝口哲也

［研究の目的］

最近の日本では1年の間に，介護が必要となる人は40万人ずつ増えている。この中に高齢者が多く，厚生労働省は「認知症高齢者の日常生活自立度」Ⅱ以上の高齢者の数は470万」という予想を発表しており，10年後の将来には日本の人口の4人に1人は高齢者と予想されている。もちろん，全員が介護を必要とするわけではないが，仕事をできる人がとても少ないということになり，介護サービスと従業者不足が問題となる。看護の効率化と二十四時間看護の提供が可能な体制を構築するために，介護者を補助して，要介護者の感情認識ができる人工知能設備を提案する。本研究では，人間の力なしに機械が自動的に要介護者の音声と顔データから特徴を抽出してくれるディープニューラルネットワーク（DNN）と畳みこみニューラルネットワーク（CNN）を用いて学習して，要介護者の顔検出と感情認識するシステムを開発する。本研究の成果により，リアルタイムに要介護者の感情状態を認識し，介護者の介護を補助して介護効率を上げることに貢献することが期待できる。

［研究の内容，成果］

1. 研究背景

2017年，日本では介護が必要となる人は40万人ずつ増えている。この中に高齢者が多く，厚生労働省は「認知症高齢者の日常生活自立度Ⅱ以上の高齢者の数は470万」という予想を発表しており，10年後の将来には日本の人口の4人に1人が高齢者と予想されている。もちろん，全員が介護を必要とするわけではないが，高齢者が増えることにより，労働力が不足し，介護サービスと労働者不足が問題となる。介護事業中における，認知症高齢者の介護のポイントは，要介護者は自分が必要な存在だと認識させることである。本人ができることと感情状態は何かを把握して，できることをお願いすると，達成感や互いの信頼感につながる。そうしたら，介護上の負担を軽減するためには，要介護者の感情認識を支持する援助が重要である。

感情認識は社会的コミュニケーションの基礎となる能力であり，心理学の領域では，人間には6つの基本的な感情がある。それは怒り，恐怖，驚き，嫌悪，快楽，悲しみである。感情はさまざまな方法で伝えることができ，その中でも顔の表情は最もはっきりした感情信号である。しかしながら，これまでの研究により，高齢者

と若年者の表情認識を比較した結果，高齢者では恐怖，悲しみ，怒りなどのマイナス感情認識における感度が低いことがわかっている。従って，表情だけに頼ると，高齢者のマイナス感情認識の困難度が高い。

2. 従来研究とその問題点

従来の感情認識研究は主に表情データを用いる。近年，deep learning の研究が盛んである。そのうち感情認識に焦点を当て，一般的には，シチュエーションでの感情認識，Emotion Recognition in the Wild Challenge というコンペティションといった二つがある。MIT（マサチューセッツ工科大学）の Media Lab もコンピュータービジョンとディープラーニングを活用し，デジタル映像に映った人物の表情から感情を読み取る技術を開発している。しかしながら，要介護者の感情認識タスクは表情情報の感情認識が不十分であり，音声の感情認識も必要と考えられる。一方，音声の感情認識研究は十分に行われておらず，現状において，実用レベルの感情音声認識はまだ実現されていない。また，ディープニューラルネットワーク（DNN）の手法は大量な感情音声データ量を学習する必要がある。感情音声データの収録は難しく，学習データ数が少ないために，音声の感情認識の用いては難しい。

音声と表情のニューラルネットワークベースのマルチメディア特徴量を用いることで，高齢者の感情認識を行うことができる。本研究では，以下の問題点を解決する研究を遂行する。

学習データ量の問題：一般に大量データを必要とするニューラルネットワークベースの手法は，少量データしか用意できない感情認識において，その取扱いは難しく，拡張性に乏しい。そこで，よりコンパクト認識モデルとデータの増加手法の両方に考慮する必要がある。

感情の判定が難しい：同一表情な場合に，感情の判定結果も様々な可能性がある。例えば，悲しい話をしているときに笑顔が現れた人がいる。これは受け入れられていない事象に対する強いストレスにより，防衛反応で本来の感情とは逆の表情が表出したと検証できたる。このような場合には，単一な情報，例えば顔または音声情報を利用して感情認識を実行するのは不十分であり，マルチメディアな情報による補正が必要だと考えられる。

以上 2 点から，介護者が実生活において感情認識を用いるためには，現行の感情認識の更なる改善が必要である。そこで，申請者は以下の 2 つの新たな枠組みを含んだマルチメディア感情認識手法を提案する。

A．敵対性学習を用いたニューラルネットワークを用いて平静音声を感情音声に変換する手法が提案されている。GAN は少量の人間の感情音声データを用いてモデルを学習し，人間と似ている感情音声を生成して，感情音声のデータ量を増加する。

B．単一の情報の認識では不十分である。そこで，マルチメディア情報（音声と表情画像）を用いて，両方の重みを自動的に考慮できたハイスピード畳み込みニューラルネットワーク（H-CNN）認識モデルを提案し，高精度かつ高速な感情認識を実現する。

3. 研究内容

認知症高齢者の感情認識率の精度を上げることを目標として，音声と表情の特徴量を用いることで，ハイスピードなマルチメディア感情認識システムを開発することが目的である。本研究の目的を達成するための提案システムの全体像を図 1 に示す。提案システムは，① 音声の感情変換と認識システム，② 顔の表情認識システム，③ マルチメディア感情認識システムの 3 つのシステムから構成される。

①　感情音声変換と認識システム

感情音声変換の流れはデータの学習とデータの変換，二つの部分がある。まず，データの学習階段で，入力の感情音声と出力の目標感情音声を変換モデルに入れて学習し，変換の教師関

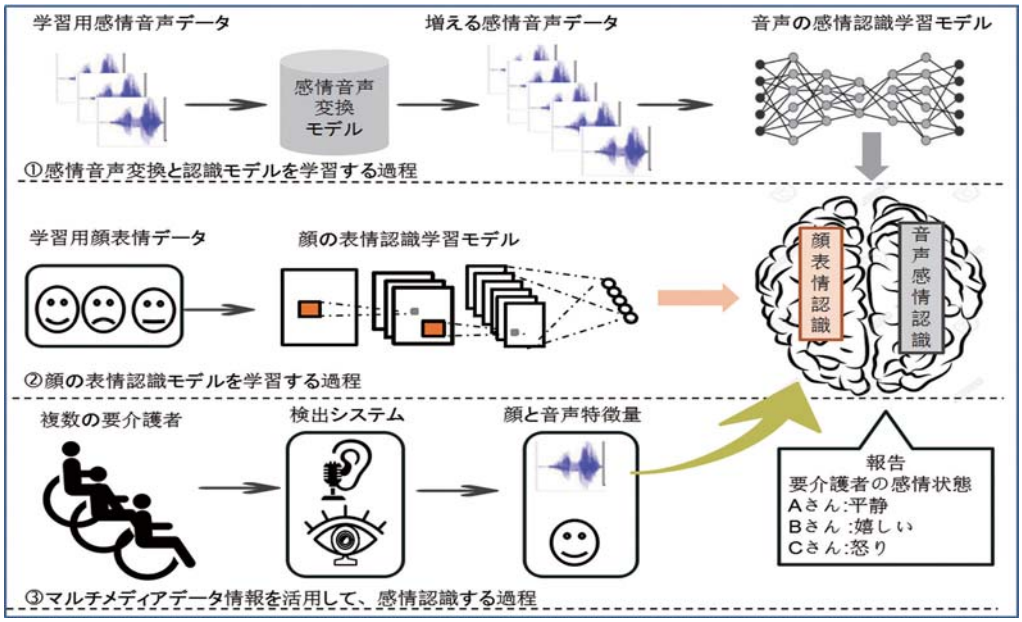

図 1

数を推定する。この教師関数を通じて，特定の目標感情音声に変換できる。感情変換モデルにより，既存の感情音声データを新しい感情音声に変換することができる。感情音声変換モデルの変換精度をあげると，より人間のような感情音声合成を作成することが可能となる。増加された感情音声のデータを用いて，敵対性学習を用いたニューラル（GANs）モデルを提案する[1]。この増加された感情音声データは次に感情音声認識モデルの学習データとして使われ，音声の感情を認識できる関数が推定される。

② 顔の表情認識システム

画像認識分野において，CNN は多くのタスクで驚異的な性能を達成し，注目を集めている。しかしながら，認識率はまだ生活に応用できるレベルに到達していない。認識精度と計算スピードをあげるために，本研究では新たに要介護者に適用するために高速化された CNN モデルを提案する。表情認識モデルの学習に，人が感情表現を行っている短い動画データベースを使っている。この動画から顔の表情画像を用いて CNN モデルを学習する。

③ マルチメディア感情認識システム

マルチメディア感情認識システムを用いて，要介護者の顔表情と音声を検出して，顔の表情と音声の韻律情報を統合して，過程①にて学習した音声の感情認識モデルと②にて学習した顔表情認識モデルを用いて，要介護の高齢者の感情を認識する。新しく提案するマルチメディア感情認識手法は，新しい音声認識モデル GAN と高速化された表情認識モデル H-CNN を互いに補助できるように組み合わせて，認識した結果を介護者のところへ送り，介護者は要介護者の感情状態を 24 時間介護できるようになる。

4. 特色と独創的な点

ニューラルネットワークの機械学習が人間の脳のように，潜在的で複雑な特徴量を学習可能となる。非線形感情認識タスクにとって，良い効果が得られるが，ネットワークの学習に大量なデータが必要である。

a．本研究は使用しているモデルは 7 層のディープニューラルネットワークである。感情音声変換の手法を応用した，少量の感情音声でも感情認識ができるモデルを提案した。感情音声変換の流れはデータの学習と変換という二つの部分がある。まず，データの学習階段では，入力感情音声と目標感情音声を用いて 7 層のモ

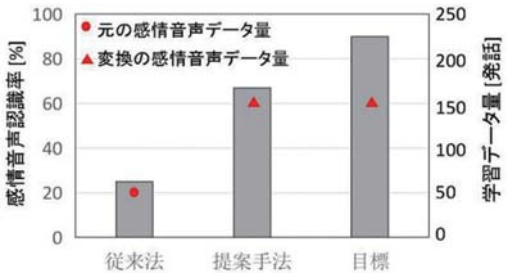

図 2

デルを学習し，変換関数を得る。そして，得た変換関数を通じて，特定の感情音声変換ができる。収録が難しい感情音声データを平静音声から変換することで，感情認識用の学習データを増加させることができ，音声の感情認識精度の向上が得られた（図 2：提案手法の変換により，データ量を三倍に増加させた[2][3]）。マルチメディアのデータを活用して，さらに 25% の感情認識精度向上を行う。

b．少量感情音声データを用いた感情認識モデリング。

感情データの収集が難しいが，これまでのディープニューラルネットワークの学習には，膨大な学習データが必要としていた。本研究の特色は，少量の感情音声データで，敵対的生成ネットワークに基づいて，感情音声変換のモデルを使う。その上で，機械で自動的に感情音声のデータを生成し，感情データの不足問題を解決する。敵対的生成ネットワークは生成モデル（Generation Model）と判別モデル（Discriminative Model）の相互博戦学習によってかなり良い結果を生む。GAN がディープラーニングに必要なデータの量を大幅に増加させることができる。

c．マルチメディアデータ情報を活用して，感情認識の精度を上げる。

従来の感情認識は主に顔，あるいは音声だけの認識だが，人間が他者の感情を判断する際，単一の情報だけではなく，顔の表情と音声の韻律情報の両方を用いる。本研究はマルチメディアデータ情報を用い，機械学習に基づき情報統合を実現し，感情認識の精度を改善する。

d．要介護の生活の質の向上

本研究が実現すれば，アプリケーション応用において，マルチメディア情報を使い，高齢者のマイナスの感情を認識することが可能になる。少ない感情音声の情報で感情認識が実現できれば，感情音声データ収集の問題を解決することができる。このアプリケーションはリアルタイムに要介護者の感情状態を認識し，看護の効率化と二十四時間看護の提供が可能になる。また，少子高齢化の進む国において，高齢者とのコミュニケーション支援にも繋がる技術である。彼らの生活品質の向上に役に立つものだと考えられる。

［今後の研究の方向，課題］

深層学習発展につれ，スマートフォンの siri や接客ロボット eco に代表されるような対話ロボットが広く応用されることとなった。しかし，対話ロボットは実践において多くの不足が指摘されている。例えば，アクセントや方言，入力されうる音声の多様性や集音環境によるの問題，高精度な音声認識を達成していない。このような限は，現場の応用においてクライアントの満足度を下げ，介護，受付，給仕などより専門性を要する現場では，応用が困難となる。また，深層学習には，大量の学習データが必要ですが，データ量が足りない場合にモデルの汎用性が低いことになっている。でも多いの場合に，大規模な学習データが収集されることは困難です，そのために，多くの受付や介護の仕事はまだ人間スタッフが担当している。しかし，高齢化の進展につれ，労働力の不足がいずれ到来する課題である。対話システムロボットの応用は労働力の不足を解決する重要なカギである。ロボットの投入は，人間スタッフが行う作業を補うだけではなく，感情，経験などの要素を排除し，質の安定したサービスを提供するという大きなメリットがある。人間とロボットが作業においてそれぞれメリットとデメリットがある。本研

究は人間とロボットそれぞれの長所を生かして「人間とロボットの共同建築対話システム」を課題として研究になる。本研究課題の本質的な学術的「問い」は，人間とロボットの共同建築システムはどのように創出されるかである。本研究のように，一方，人間の柔軟性と自然性を利用し，知能対話システムでは解決ができない問題を補完する，他方，対話システムが有する音声認識（ASR）と自然言語処理（NLP）を利用し，基本対話の問題を解決して作業の効率性，人間がもつ記憶の限界などに対処する。さらに，感情認識と感情音声変換などの技術を対話システムに実装されて，客や非介護者の感情を認識し，人間スタッフに適切な回答内容を勧める。感情音声変換がサービスの需要に応じて，ロボットは作業者の対応音声を機能付き感情音声（ハイテンション，説得力）を変換し，作業者の表現力を強化できる。

［成果の発表，論文等］

［1］ Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, Yasuo Ariki : “Emotional Voice Conversion Using Dual Supervised Adversarial Networks With Continuous Wavelet Transform F0 Features”, IEEE-ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING 27 (10) 1535-1548 10, 2019.

［2］ Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, Yasuo Ariki : “Neutral-to-emotional voice conversion with cross-wavelet transform F0 using generative adversarial networks”, APSIPA TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING, 8 3, 2019.

［3］ Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki : “Speech Prosody Conversion using Sequence Generative Adversarial Nets with Continuous Wavelet Transform F0 features”, 日本音響学会 2019 年春季研究発表会講演論文集 1125-1128 3 2019.