様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ５ 日現在

機関番号：３４３１５
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１８９８

研究課題名（和文）音声における感情を表現する特徴量の抽出に基づいた感情音声

研究課題名（英文）Speech Emotion Recognition Based on Extracting Features for Emotion Expression

研究代表者
山下　洋一（Yamashita, Yoichi）

立命館大学・情報理工学部・教授

研究者番号：８０１７４６８９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：音声は言語情報だけでなく、感情などのパラ言語情報も伝達する。人同士の対話では、そのようなパラ言語情報も自然に利用することでスムーズな対話が実現されている。人と機械の間での音声による円滑な情報交換を実現するには、感情の認識が重要な枠割を果たす。音声における感情認識の性能向上を実現するために、(1) 音響的特徴と言語的特徴を併用した音声感情認識、(2) ソフトラベルを利用した感情認識モデルの学習、(3) 話題を利用した音声感情認識、(4) 音声の短区間を対象とした音声感情認識、の課題に対して、新たな手法を提案し、各手法の有効性を確認した。

研究成果の学術的意義や社会的意義
音声認識技術の急速な進歩によって、音声から文字テキストへの変換は実用化されたと言ってよい状況になっている一方で、音声が伝達する情報のうち、感情など言語情報以外の情報の自動認識はまだ発展途上にある。本研究では、音声における感情認識の性能を向上させるための手法を開発した。また、音声においては、一発話の途中で感情が変化することは一般的に起こりうることであり、音声における短い単位に対して感情を推定する手法の開発が求められており、音声の短区間を対象とする感情認識の手法についても取り組んだ。

研究成果の概要（英文）：Speech conveys not only verbal information but also paralinguistic information such as emotions. In conversations between people, such paralinguistic information is used naturally to achieve smooth dialogue. Emotion recognition plays an important role in realizing smooth information exchange by speech between humans and machines. In order to improve the performance of speech emotion recognition, we proposed new methods and confirmed their effectiveness as follows; (1) speech emotion recognition using both acoustic and linguistic features, (2) emotion recognition model training using soft labels, (3) speech emotion recognition using topics, and (4) speech emotion recognition for short segments of speech.

研究分野：音声情報処理

キーワード：音声感情認識　言語情報　ソフトラベル　話題　短区間　CTC

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

音声は人がコミュニケーションを行うための自然かつ容易なチャンネルであり、音声による機械と人との情報交換を実現するために、音声認識・音声合成の研究が精力的に行われてきた。音声は、文字テキストで記述される言語情報だけでなく、感情、意図、年齢、性別といった音声の書き起こしテキストでは記述されないパラ言語・非言語情報も伝達する。音声合成では、統計的手法に基づいた音声合成によって感情や個人性を制御することが可能になり、様々な感情での音声合成が実用化されている他、多様な声質の音声合成システムも販売されている。また、音声から言語情報を取り出して音声を文字テキストに変換する、いわゆる音声認識でも、HMM (Hidden Markov Model) や DNN (Deep Neural Network) などの統計的手法の導入により、高い音声認識率が達成され「しゃべってコンシェル」「Siri」などの音声認識サービスが実用化された。しかし、これらのサービスでは、言語情報のみが認識の対象となっており、人と会話するように情報交換を行うことのできる音声対話システムを実現するためには、言語情報だけでなくパラ言語・非言語情報の認識が期待されている。

２．研究の目的

音声における感情認識では、発話あるいは発話を分割した短区間を対象として、感情認識が行われる。本研究では、発話や短区間における感情性を表現する特徴パラメータを抽出し、音声における感情を精度良く自動認識する手法を開発することを目的とした。

３．研究の方法

(1) 音響的特徴と言語的特徴を併用した音声感情認識

従来の音声感情認識の課題として、音響的な情報と言語的な情報を統合する手法の分析が不十分であることが挙げられる。人は、他者の音声を聞いて、相手の感情を予測するとき、声のトーンの高低や声量、声質などを手掛かりにする。また、話している単語や文も同様に、感情を決める手掛かりになる。よって、人は、音声から感情を予測するとき、音響的な情報と言語的な情報から感情を決定していると考えられる。音響情報からの判断と言語情報からの判断は、それぞれ、音声感情認識とテキスト感情認識に置き換えることができる。音声感情認識・テキスト感情認識の予測結果と、各認識器から得られた総合的な情報からの予測結果から、最終的な感情を正確に予測することができれば、人間のような音声からの感情の認識が可能になることが期待できる。しかし、音声感情認識とテキスト感情認識を網羅的に組み合わせ、最も適した統合方法の分析や、情報統合による予測結果の分析をする研究は不足している。以上の課題を解決するために、音声感情認識とテキスト感情認識から得られる、中間的特徴と予測結果を組み合わせた感情認識手法を提案した。中間的特徴とは、ネットワークの出力層よりも手前の中間層から得られる特徴を指す。中間的な特徴を融合する early fusion と、予測結果を融合する late fusion の利点を生かし、既存の感情認識器よりも正確に感情を認識することを目指した。

(2) ソフトラベルを利用した感情認識モデルの学習

一般的に、ニュートラル感情は他の感情と比較して感情的に中立な状態であるとされている。一方で、音声感情認識の従来研究では、ニュートラル感情が感情クラスとして独立していないことが指摘されている。感情クラスが独立していない原因として、ニュートラル感情とされる発話が他の感情の特徴を持つことが挙げられる。ニュートラル感情は曖昧であり、これがニュートラル感情の特性とも言える。これまで、感情の曖昧性を考慮した研究がいくつか提案されている。一般的に、機械学習に基づく音声感情認識では、発話をただ一つの感情クラスに割り当て、同じクラスに属する発話は類似する音響特徴を持つことを前提とした手法が多く提案されている。これらの研究では、複数の評価者による多数決で決まるただ一つの感情を正解ラベルとして学習に利用する。この正解ラベルはハードラベルと呼ばれている。しかしながら、感情の知覚は人によって様々であるため、感情には曖昧さを伴うことがあり、ハードラベルでは感情の特性を充分に表現できているとは言えない。この問題を解決するために、従来研究ではソフトラベルを用いた手法が提案されている。ソフトラベルは分布表現によって感情の曖昧性を表現できる。これらの手法では、感情クラスの中でも特に曖昧性の高いニュートラル感情に着目しておらず、他の感情クラスと同様にニュートラル感情を扱っている。このため、ニュートラルの曖昧性を正確にモデル化しているとは言えない。

本研究では、ニュートラル感情の特性を考慮して学習を行う手法を提案した。具体的には、感情の曖昧性を考慮した従来手法を基に、ニュートラル感情の曖昧性をソフトラベルによって考慮する音声感情認識モデルを構築した。識別境界に近い発話の事後確率は一様分布に近く、

最も尤度の高いクラスは非常に不安定である。ソフトラベルを用いて感情認識のモデル化を行うことで、識別境界に近い発話が持つ感情の特徴を上手く学習できる。その結果、類似するクラス間の識別誤りが低減され、最も尤度の高いクラスが安定して正解となる。ニュートラル感情の曖昧性を考慮することで、より精度の高いモデル化を実現できる。

(3) 話題を利用した音声感情認識

感情の表現や知覚にとって、性格や話題などのコンテキストは重要である。コンテキストには様々な要因があり、ミクロレベルの個人的な要因から、マクロレベルの社会的な要因まで多岐に渡る。Greenaway らはコンテキストを個人的要因、状況的要因、文化的要因の 3 つに分類している。個人的要因は、年齢や性別、性格など、状況的要因は場所や人間関係、背景など、文化的要因は社会構造や言語などを含む。また、状況的要因は、時間的に短期的要因から長期的要因まで広がるが、感情を認識する上では、より長期的な要因にも注目する必要がある。例えば、親しい人が亡くなったことを他人に話す際、悲しみの感情が多く観測され、喜びや怒りなどは観測されにくいと考えられる。親しい人の死という話題が悲しみの感情を引き出しているといえる。この例のように、話題によって観測される感情がある程度決定する可能性がある。

認知心理学の表情認識の分野では、話題に着目した研究が進んでいる。Carroll らは表情から得られる情報より、話題に沿って感情を判断する傾向があると述べている。彼らが行った実験では、評価者に怒りを誘発する話を聞かせた後に恐怖の表情を提示すると、事前に話を聞かせない場合に比べ、その表情を怒っていると誤って判断する結果を示した。また、笑顔の意味が話題によって変化することを主観評価実験によって示した研究もある。さらに近年では、機械学習ベースの音声感情認識において、話題の変化が認識性能に与える影響を分析した研究が発表された。この研究では、学習セットと評価セットに異なる話題の発話を用いた場合、同じ話題の発話を用いた場合に比べて、性能が劣ることを示した。以上のことから、音声感情認識においても話題と感情は密接に関係していると考えられる。本研究では、音声感情認識において、話題情報を利用することで感情認識の性能改善を目指す。

(4) 音声の短区間を対象とした音声感情認識

音声感情認識では、一般に、発話を単位とした感情認識が行われる。この場合、入力音声の長さに依らず、時間の経過とともに変化する感情を逐次的に推定できないため、応答性が必要な感情解析には応用できない。このため、発話単位ではなく、短い発話を単位とした音声感情認識の研究にも取り組む。本研究では、発話を分割した数フレームで構成される短区間に対して、音声感情認識を実現する手法を提案する。提案する手法では、文字に感情の情報を付与したラベル列（以降、文字付き感情ラベル列）を入力音声から推定する。音声の数フレーム単位に対応する文字を推定する枠組みを基に、感情の情報が付与された文字を推定することで、数フレーム単位に対応する感情の推定とその評価を可能にする。また、文字付き感情ラベル列の推定と合わせて文字単位の音声認識を同時に行うマルチタスク学習を用いた手法も提案する。従来研究において、音声感情認識と音声認識を同時に最適化することで、感情の認識率が向上することが確認されている。同様に、文字付き感情ラベル列の推定と音声認識を同時に最適化することで、認識性能が向上することが期待できる。

４．研究成果

(1) 音響的特徴と言語的特徴を併用した音声感情認識

感情認識の事前学習段階では、音響的特徴を用いる音声感情認識と言語的特徴を用いるテキスト感情認識をそれぞれ個別に学習する。音声感情認識では、 CNN と BLSTM、Attention を組み合わせた構造を採用した。ネットワークの入力部では、複数フレームからなる音響特徴量から、数フレームのセグメント特徴量を抽出し、入力を繰り返す。入力を、発話全体ではなくセグメント単位にすることで、長さの異なる音声データへの対応も容易になる。ネットワークへ入力する音響特徴量はメルケプストラム、出力は 4 感情（喜び、悲しみ、怒り、平静）の予測結果である。テキスト感情認識では、BERT を採用した。ネットワークへの入力はテキスト、出力は 4 感情（喜び、悲しみ、怒り、平静）の予測結果である。ここで、感情認識の性能向上に焦点を当てるため、音声認識結果のテキストではなく、データセットに記録されている読み上げ用のテキストを使用した。

音声・テキスト感情認識から得られる情報を利用した感情認識の学習段階では、事前に学習した音声感情認識とテキスト感情認識から得られる中間的な特徴と予測結果を利用した融合モデルを学習する。early fusion と late fusion を組み合わせた。Early fusion は、事前に学習させた各認識器から得られる中間的な特徴を融合する。Late fusion は、事前に学習させた各認識器から得られる予測結果や、中間的な特徴から得られた予測結果を融合する。二つの fusion を組み合わせることで、それぞれの fusion の利点を生かした学習が可能になることが期待できる。

提案手法の評価では、日本語感情音声データ JTES (Japanese twitter based emotional speech)[1] を使用した。JTES は、武石らによって構築された感情音声データセットの一つであり、日本語音声感情認識の研究によく利用されている。このデータセットは、音素と韻律の

バランスを考慮して作成されており、4 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 文が、男性 50 名、女性 50 名の計 100 名によって読み上げられている。音声感情認識の評価では、データセットを 5 つに分割し、話者の重複を許さない（話者オープンな）cross validation で行った。また、テキスト感情認識の評価も同様に、データセットを 5 つに分割し、テキストの重複を許さない（テキストオープンな）cross validation で行った。音響・言語情報を利用した音声感情認識の評価は、話者とテキストの重複を許さない 5-fold cross validation で行った。

感情認識の結果は、認識クラスの出現数を重みとして考慮する平均認識率 WA (Weighted Accuracy) と考慮しない平均認識率 UA (Unweighted Accuracy) で評価した。音響情報のみまたは言語情報のみを用いた感情認識の WA と UA は、約 66%～71%であったが、音響・言語情報を用いた感情認識では、14%～21%の性能改善が見られた。

(2) ソフトラベルを利用した感情認識モデルの学習

ニュートラルの特性を考慮したソフトラベルによるラベル平滑化手法として 2 種類の手法を提案した。提案手法 1 では、ニュートラル感情とされる発話が他の感情の特徴を含むことを考慮する。具体的には、発話の正解がニュートラル感情の場合には、その他の感情にも等しく値を持つようにラベルの平滑化を行う。発話の正解がニュートラル感情でない場合には、ハードラベルを利用する。提案手法 2 では、ニュートラルとされる発話が他の感情の特徴を含むならば、ニュートラルの発話とその他の発話が類似する特徴を持つと考えられる。つまり、ニュートラル以外の感情の発話がニュートラルの特徴を持つ可能性がある。これを考慮したソフトラベルを提案する。具体的には、発話の正解がニュートラル感情の場合には、提案手法 1 と同様に、その他の感情にも等しく値を持つように平滑化を行う。発話の正解がニュートラル感情でない場合には、正解感情とニュートラル感情に値を持つように平滑化を行う。

評価実験には、Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)[2] を用いた。IEMOCAP は、対話を収録したデータセットであり、音声だけでなく表情やジェスチャーも収録されている。男女各 5 名の俳優、計 10 名の話者による演技対話と即興対話の 2 種類が提供されている。各対話は 5 つのセッションで構成され、1 つのセッションには男女ペアの対話が収録されている。約 12 時間 33 分の 10,039 発話が存在する。本研究では、日常会話の音声感情認識を想定し、即興対話の発話のみを用いた。IEMOCAP は、Happiness、Anger、Sadness を含む 10 種類の感情ラベルを採用している。このラベルを用いて、3 名の評価者が各発話に感情ラベルの付与を行っており、評価者の多数決で決まる感情を、各発話の正解クラスとして用いた。

提案手法は、従来手法と比較して手法 1、手法 2 とも UA では大幅な性能の変化はみられなかったが、WA では提案手法 2 のソフトラベルで精度の向上がみられた。特に、従来手法と比較して Neutral の精度が向上した。Neutral は最も発話数の多い感情であり、Neutral の結果が WA の結果に支配的になっているためである。提案手法 1 の各感情の結果では、Happiness、Anger、Sadness の性能が向上した。特に Happiness は 4.01%と顕著な向上がみられた。

(3) 話題を利用した音声感情認識

音声感情認識において、話題情報を利用する 2 種類の手法を提案した。音声感情認識は、深層学習を用いて行われ、モデル学習時に話題情報を与える。提案手法 1 では、話題情報を one-hot 表現の話題ラベルとして与え、提案手法 2 では、任意の話題は各感情の出現割合で表現できると考え、話題が持つ感情の出現割合を話題情報として与える。

評価は IEMOCAP を用いて行った。IEMOCAP では、友人の死や別れなど誰もが経験しうる状況を元に 8 種類のシナリオが設定されており、そのシナリオに基づいたやり取りが収録されている。本研究では、このシナリオを話題として扱った。

提案手法 1 では、話題情報を利用することで話題情報を利用しない場合と比べて、WA と UA がそれぞれ 1.9%、1.44%向上した。提案手法 2 では、話題情報を利用することで WA と UA がそれぞれ 1.12%、1.03%向上した。これらの結果から、音声感情認識に話題情報を利用することの有効性が示された。

(4) 音声の短区間を対象とした音声感情認識

提案手法は、入力音声から文字付き感情ラベル列を認識する短区間音声感情認を CTC (Connectionist Temporal Classification) モデルに基づいて実現される。短区間は、文字に対応する発の数フレームを示す。数フレームの入力音声を「喜び」、「悲しみ」、「怒り」、「平静」の 4 感情カテゴリに分類する。音声認識と同様に、入力音声の数フレームに対応する文字付きの感情ラベルを推定することで、数フレーム単位で音声からの感情を認識する。CTC とは、時系列のパターンを分類する手法の一つである。この手法では、blank 記号の導入とシンボルの繰り返しを許容することによって、入力系列長以下の出力系列の推定をニューラルネットワークで学習できる。CTC モデルを利用することで、入力音声の各フレームに対する感情ラベルを用意せずに、数フレーム単位の感情を認識できるようになる。文字付き感情ラベル列は、発話内容を示すテキストの各文字に感情カテゴリを示すインデックスを付与した記号列である。例えば、"EMOTION" という文字列に "2 (happy)" という感情ラベルが付与されていた場合、文字

付き感情ラベル列は、['2E','2M','20','2T','2I','20','2N'] となる。

評価は IEMOCAP を用いて行った。即興対話と演技対話の音声を合わせて使用し、音声を「喜び」「悲しみ」「怒り」「平静」の 4 感情に分類する認識器の学習、検証、評価を行った。なお、「喜び」のデータには「驚き」のデータを追加した。また、データ不足を緩和するために、話速変換によるデータ拡張を行った。話速変換には、waveform similarity based overlap add (WSOLA) アルゴリズムを用い、元の発話の話速を 0.9 倍、1.1 倍にし、音声データ数が 3 倍になるよう拡張した。最終的に実験に使用した発話数は、5,531×3 = 16,593 文である。

文字付き感情ラベル列の推定のみの学習を行った場合（single-task）と音声認識とのマルチタスク学習を行った場合（multi-task）について、各手法での置換誤り、削除誤り、挿入誤りの内訳と CER と ESER、ECER で評価する。文字付き感情ラベル列の推定のみの結果では、CER が約 19%、ECER が約 53%となった。特に、ESER は約 47%となっており、提案手法では約 50%以上は数フレーム単位で感情を正しく推定できると言える。単一タスク学習での結果とマルチタスク学習での結果を比較すると、マルチタスク学習によって文字推定精度は向上したものの、感情ラベル列と文字付き感情ラベル列の推定精度は低下した。主タスクである文字付き感情ラベル列の推定と類似する音声認識を補助タスクとして同時に最適化することで、音声認識が安定して学習できた一方で、文字付き感情ラベル列の推定が十分に学習できなかった可能性がある。したがって、主タスクの推定精度を向上させるためには、感情ラベル列の推定または主タスクとは異なる感情に関連する情報の推定などを補助タスクにする必要があると考えられる。また、発話を単位とした感情認識の精度では、提案手法の結果は従来手法の結果を上回った。数フレーム単位で感情を認識する提案手法は、発話単位においても従来手法と同等以上の認識性能があることを確認した。

＜参考文献＞

[1] E. Takeishi, T. Nose, Y. Chiba, and A. Ito: "Construction and analysis of phonetically and prosodically balanced emotional speech database," in 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), pp.16-21, 2016.
[2] 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.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 秋山大知，石川智希 井本桂右，新妻雅弘，山西良典，山下洋一	４．巻 76
２．論文標題 音声を用いた感情認識のための学習話者の選択	５．発行年 2020年
３．雑誌名 日本音響学会誌	６．最初と最後の頁 554-561
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Ryotaro Nagase, Takahiro Fukumori, Yoichi Yamashita
２．発表標題 Speech Emotion Recognition Using Label Smoothing Based on Neutral and Anger Characteristics
３．学会等名 2022 IEEE 4th Global Conference on Life Sciences and Technologies（国際学会）
４．発表年 2022年

１．発表者名 永瀬亮太郎，福森隆寛，山下洋一
２．発表標題 「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
３．学会等名 日本音響学会2021年秋季研究発表会
４．発表年 2021年

１．発表者名 永瀬亮太郎，福森隆寛，山下洋一
２．発表標題 音響情報と言語情報を利用した短区間の音声感情認識
３．学会等名 日本音響学会2022年春季研究発表会
４．発表年 2022年

１．発表者名
大澤まゆ子，井本桂右，山西良典，福森隆寛，山下洋一

２．発表標題
ニュートラル音声の特性を考慮したソフトラベルによる感情分類モデル学習

３．学会等名
日本音響学会2020年秋季研究発表会

４．発表年
2020年

１．発表者名
永瀬 亮太郎，福森 隆寛，山下 洋一

２．発表標題
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

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

１．発表者名
永瀬亮太郎，福森隆寛，山下洋一

２．発表標題
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

３．学会等名
電子情報通信学会技術研究報告

４．発表年
2020年

１．発表者名
R.Nagase, T.Fukumori and Y.Yamashita

２．発表標題
Speech Emotion Recognition with Fusion of Acoustic- and Linguistic-Feature-Based Decisions

３．学会等名
APSIPA Annual Summit and Conference 2021（国際学会）

４．発表年
2021年

１．発表者名 永瀬亮太郎，福森隆寛，山下洋一
２．発表標題 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
３．学会等名 日本音響学会2022年秋季研究発表会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関