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　４ 年　 ５ 月 ２０ 日現在

機関番号：１１５０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２０１４

研究課題名（和文）高精度な感情音声認識技術を用いた音声からの感情推定の研究

研究課題名（英文）Emotion detection from speech using high-accuracy emotional speech recognition

研究代表者
小坂　哲夫（Kosaka, Tetsuo）

山形大学・大学院理工学研究科・教授

研究者番号：５０３５９５６９

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：音声を対象とした感情推定の研究は深層学習の利用により大きく進展している．しかし従来は音声に含まれる非言語情報を利用した方法が主流であり，言語情報を利用した方法の検討は不十分であった．本課題では言語情報を利用するため，音声認識により得られた発話内容テキストを利用し感情認識を行った．さらにそれを非言語情報による感情認識と融合し，高精度な感情認識を実現した．最終的には4感情の認識で82.75%の認識率を得た．

研究成果の学術的意義や社会的意義
音声を対象とした感情認識の研究において，これまでも言語情報と非言語情報を融合した感情認識の検討は様々行われてきたが，最大の問題点は音声認識の性能の低さであった．この性能が低いと正しい言語情報が利用できず，これが融合法のネックとなっていた．本研究ではその問題を克服し，高精度の感情認識を実現した．
音声による感情認識の精度が向上すれば，人間同士の対話に近い人対機械の対話が可能になり，ロボットの活用の幅が広がると考えられる．さらに本技術はコールセンターやメン
タルヘルスケアなど音声を使う様々な分野で利用可能と考えられる．

研究成果の概要（英文）：Research on speech emotion recognition has made great progress by using deep learning. However, the methods using non-verbal information contained in speech were the mainstream in the past, and the study of the methods using linguistic information were insufficient. In this project, in order to use linguistic information, emotion recognition was performed using the texts obtained by speech recognition. In addition, the above recognition results were fused with the results by non-verbal information to realize highly accurate emotion recognition. Finally, the recognition rate of 82.75% was obtained by recognizing four emotion categories.

研究分野：音声情報処理

キーワード：感情認識　音声対話システム　深層学習　言語情報　非言語情報　音声認識

１．研究開始当初の背景

以前より機械と人間が音声による自然言語でコミュニケーションを行なう音声対話システムの研究が行なわれている．情報検索など，何らかの目的を持って対話を行なうタスク指向型対話システムが永らく検討される一方，近年では目的を持たず雑談的に対話を行なう，非タスク指向型のシステムも検討されている．このようなシステムの場合，人対人のコミュニケーションと同様の対話が出来ることが重要である．人間同士の対話においては言語情報のみならず非言語情報も重要な役割を果たす．Birdwhistell によれば会話でやりとりされるメッセージのうち言語そのものによる情報は全体の 30～35%で，残りは非言語によって伝えられるとしている[1]．非言語情報は表情や視線のような視覚的に得る情報もあるが，音声に含まれる非言語情報も重要である．例えば，声の高さ，大きさ，時間長の他咳払いや溜息なども含まれる．本研究では，この音声に注目して非言語情報である感情の認識を検討する．

音声に含まれる感情を認識する研究は以前より行なわれている．典型的な感情認識の手法について概説する．まず基本的な特徴として音声から Low-Level Descriptors (LLD) が抽出される．この特徴時系列に対し発声全体から各種の統計量を計算し，対象とする発話の特徴量とする．得られた特徴量を統計的識別器，たとえば Support Vector Machine などを利用して感情の分類を行なう．以上の方法では音声の音響的な特徴のみを使用している．一方発話内容からでも感情を推定できる場合がある．「負けて悔しい」，「映画が楽しかった」など直接感情を言葉で表わしている場合もある．このため発話内容も併用して感情推定する研究も行なわれている．この方法を採用するためには，音声認識技術を用いて音声をテキストに変換する必要があるが，そもそも感情を含んだ音声の認識(感情音声認識) が難しいという問題点がある．

２．研究の目的

本研究では言語情報と音響情報を統合した感情認識手法の確立を目的とする．両情報を統合した認識手法はこれまでも検討されているが，感情音声認識の性能が悪いという問題があった．このため誤った言語情報を使用することにより期待した効果を得るのが困難であった．以上の問題の解決のために，まず感情音声の認識性能の向上について検討を行なう．方法しては適応技術を使用し，音響モデルおよび言語モデル両方に対する同時適応を検討する．次に得られた言語情報すなわち感情音声の認識結果から感情を推定する方法ついて検討する．近年深層学習モデルを用いたテキストからの感情推定の研究が盛んに行われている．しかし認識誤りを含みかつ話し言葉によるテキストからの感情推定は殆ど検討されていない．深層学習モデルが，このような対象についても有効であるか明らかにする．最後に音響情報による感情推定の結果と言語情報による感情推定結果を統合する手法を検討し最終的な感情推定結果を得る．

３．研究の方法

提案法のシステム構成図を図 1 に示す．言語情報による推定では，音声認識により発話内容を取得し，その後感情識別器で推定結果候補を得る．また音響情報による推定では，音響特徴量を計算した後，感情識別器で推定する．最終的に結果統合部で統合され推定結果を得る．特に本研究では以下の 3 つの課題に取り組む．

課題 1　モデル適応手法を用いた感情音声認識の性能向上

課題 2　深層学習モデルを用いた音声認識結果からの感情推定

課題 3　言語的情報と音響的情報統合の手法

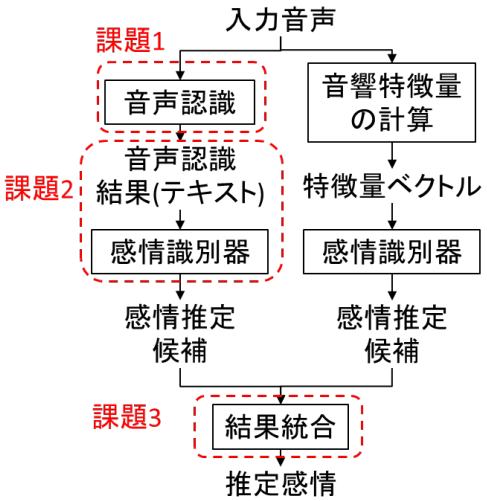

図 1　提案法のシステム構成図

課題 1 では音響モデルおよび言語モデルについての適応を行なう．音響モデルにはハイブリッド型深層学習モデルの DNN-HMM(deep neural network hidden Markov model)，言語モデルには n-gram を使用する．言語モデル適応において特に問題なのは適応データの少なさである．この問題を解決するためにツイートデータを利用する．ツイートには口語的表現や感情が含まれる表現が多く存在する．またオンライン上から大量に取得可能であり，適応データに適していると考えられる．課題 2 では BERT(Bidirectional Encoder Representations from Transformers)と呼ばれる深層学習モデルを利用する[2]．BERT は自然言語処理の分野で現在特に注目されているモデルである．しかし誤りを含んだ話し言葉を対象とした応用に関する検討は進んでいない．本研究ではそのような対象についての有効性

を明らかにする．課題 3 については 2 種類の方法が存在する．特徴量の時点で両者を統合しそれを識別器に入力する方法と，両者でまず感情推定を行いその結果を統合する方法である．予備実験の結果後者で良好な結果が得られたので，結果統合による方法について検討を進める．

４．研究成果

(1) 音響モデルおよび言語モデルの同時適応による感情音声認識

感情音声に対する音声認識性能向上を目指し，音響モデルおよび言語モデルの同時適応を行った．音響モデルとして DNN-HMM，言語モデルとして n-gram モデルを使用する．適応前モデルとして日本語話し言葉コーパス(CSJ)で学習した音響および言語モデルを用意する．使用した CSJ の学習データは学会講演を収録したものであり，感情の表出が少なく，また語彙も偏っており感情音声の認識には適さない[3]．そこで適応データに感情音声コーパス JTES(Japanese Twitter-based emotional speech)を用い[4]音響モデルについては誤差逆伝播法による適応，言語モデルについては重み付き混合法による適応を行った．言語モデル適応については適応データが不足するため，大量のツイートデータを取得して使用した．語数としては約 2586 万語となる．JTES に対する音声認識結果を表 1 に示す．単語誤り率を示しているため，値が低いほど性能が高い．結果より言語モデル適応，あるいは音響モデル適応のみでも効果はあるが，同時適応を行うことにより大幅に性能向上することが分かる．以上より提案法の有効性が示された．

表 1 JTES における感情音声認識の結果

	適応前	言語モデル適応のみ	音響モデル適応のみ	同時適応
単語誤り率	36.11%	25.68%	26.91%	17.77%

(2) BERT を用いた音声認識結果からの感情推定

上記の音声認識で得られた認識結果を用いて感情推定を行う.感情識別器としてはBERT を用いる．日本語 Wikipedia のテキストを使用して学習を行った BERT の事前学習モデルを用い[5]，JTES のテキストデータでファインチューニングを行う．得られたモデルを用いテキストを対象とした感情認識を行った．実験においては正確に認識が行われたことを仮定した単語誤り率 0%のテキストと表 1 で得られた誤り率 17.77%のテキストを比較する．実験の結果，感情認識率は前者で 62.5%，後者で 51.5%を得た．学習データが限られているため高い認識性能とは言えないが,誤りを含んだテキストを対象としても，5 割程度の精度で認識できることが明らかとなった．

表 2-1 音響特徴のみによる感情認識率 (%)

		予測			
		neu	ang	joy	sad
正解	neu	78	1	7	14
	ang	2	78	20	0
	joy	0	31	64	5
	sad	6	0	4	90

表 2-2 言語特徴のみによる感情認識率 (%)

		予測			
		neu	ang	joy	sad
正解	neu	46	10	23	21
	ang	20	52	4	24
	joy	6	12	55	27
	sad	2	20	25	53

(3) 言語情報と音響情報の統合による感情推定

上記で得られた言語情報による感情推定結果と音響情報による感情推定結果を統合し最終的な感情推定を行う.統合に当たってはそれぞれの識別器から得られたスコアを重み付き加算することで統合後のスコアを計算し，認識率を算出する．

音響情報による感情推定では,深層学習モデルである双方向 LSTM と双方向 GRU を組み合わせたモデルを用いる．いずれのモデルも時系列特徴に対して効果的であることが知られている．音響特徴のみの感情認識率は77.5%,言語特徴のみの認識率は51.5%であるのに対し，両者を融合した結果 JTES において感情認識率 82.75%を得た.なおこのコーパスに対する人間による認識精度は 75.5%と報告されている[6]．その実験では被検者に対し言語情報を無視して音響情報のみで判断するようにとの指示があり正確な比較対象とはならないが，今回開発したシステムの認識性能は，ほぼ上限に近い状況と考えられる．

それぞれの認識実験の結果得られた混同表を表 2-1～2-3 に示す．neu は平静，ang は怒り，joy は喜び，sad は悲しみを表す．例えば言語特徴のみに

表 2-3 音響特徴と言語特徴を併用した感情認識率 (%)

		予測			
		neu	ang	joy	sad
正解	neu	74	1	4	21
	ang	4	87	7	2
	joy	0	21	74	5
	sad	2	0	3	95

よる認識では ang を neu に誤認識する割合が高く，一方音響特徴では同様の誤りは少ない．両者を併用することにより，この種の誤りが低減していることが分かる．このように音響特徴と言語特徴では誤り傾向が異なっており，両者を併用することで相補的な効果が得られ，認識性能が向上していることが分かる．

以上より提案手法において言語情報と音響情報の併用は有効で，極めて高い認識性能が得られることが分かった．今後は人間対機械の対話システムなどへの活用が期待される．またコールセンターやメンタルヘルスケアなど音声を用いる様々な分野への波及が期待される．

＜引用文献＞

[1] R.L. Birdwhistell: Introduction to Kinesics, foreign Service Institute (1952).
[2] J. Devlin, et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 (2018).
[3] K.Maekawa: “Corpus of spontaneous Japanese: Its design and evaluation,” Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003, pp. 1-6.
[4] E. Takeishi, et al.: “Construction and analysis of phonetically and prosodically balanced emotional speech database,” Proc. of O-COCOSDA2016, 2016, pp. 16-21.
[5] 東北大学乾研究室, Pretrained Japanese BERT models released, https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/ (2022 年 5 月 19 日閲覧).
[6] Y. Chiba, et al.: “Multi-stream attention-based BLSTM with feature segmentation for speech emotion recognition,” Proc. of INTERSPEECH2020, 2020, pp. 3301-3305.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　4件）

１．著者名 Misaki Sakurai, Tetsuo Kosaka	４．巻 -
２．論文標題 Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results	５．発行年 2021年
３．雑誌名 Proc. of 2021 IEEE 10th Global Conference on Consumer Electronics	６．最初と最後の頁 889-892
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/GCCE53005.2021.9621810	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kazuya Saeki, Masaharu Kato, Tetsuo Kosaka	４．巻 -
２．論文標題 Language model adaptation for emotional speech recognition using Tweet data	５．発行年 2020年
３．雑誌名 Proc, APSIPA Annual Summit and Conference 2020	６．最初と最後の頁 371 - 375
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Tetsuo Kosaka, Yuka Haneda, Daisuke Makabe, Masaharu Kato	４．巻 1
２．論文標題 Investigation of acoustic models for emotion recognition using a spontaneous speech corpus	５．発行年 2019年
３．雑誌名 Proc. of the 23rd International Congress on Acoustics	６．最初と最後の頁 6795 - 6800
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 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kazuya Saeki, Masaharu Kato, Tetsuo Kosaka	４．巻 1
２．論文標題 Performance Improvement of Prosody-Controlled Voice Conversion by Language Model Adaptation	５．発行年 2019年
３．雑誌名 Proc. of the 8th global conference on consumer electronics	６．最初と最後の頁 1 - 3
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/GCCE46687.2019.9015444	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 櫻井美咲，小坂哲夫
２．発表標題 音声認識結果に基づく言語特徴と音響特徴による音声感情認識
３．学会等名 日本音響学会音声研究会
４．発表年 2021年

１．発表者名 須々田和基，櫻井美咲，小坂哲夫
２．発表標題 感情コーパスJTESを用いた時系列情報を考慮した音声感情認識
３．学会等名 情報処理学会東北支部研究会
４．発表年 2022年

１．発表者名 櫻井美咲，須々田和基，小坂哲夫
２．発表標題 音声認識結果による言語特徴と音響特徴による音声感情認識の検討
３．学会等名 日本音響学会秋季講演論文集
４．発表年 2022年

１．発表者名 羽田優花，櫻井美咲，加藤正治，小坂哲夫
２．発表標題 音声の時系列特徴量と統計量の融合による感情認識
３．学会等名 日本音響学会秋季講演論文集
４．発表年 2020年

１．発表者名
佐伯和哉，加藤正治，小坂哲夫

２．発表標題
感情音声認識を対象とした言語モデル適応の評価

３．学会等名
東北地区音響学研究会

４．発表年
2020年

１．発表者名
佐伯和哉，加藤正治，小坂哲夫，能勢隆

２．発表標題
音響・言語モデルの同時適応による感情音声認識の精度改善

３．学会等名
日本音響学会秋季講演論文集

４．発表年
2019年

１．発表者名
羽田優花，加藤正治，小坂哲夫

２．発表標題
感情音声データベースJTESを用いた音声感情認識における特徴量の検討

３．学会等名
日本音響学会秋季講演論文集

４．発表年
2019年

１．発表者名
羽田優花，加藤正治，小坂哲夫

２．発表標題
音声による感情認識における時系列特徴量と統計量の融合方法の検討

３．学会等名
情報処理学会東北支部研究会報告

４．発表年
2020年

１．発表者名
佐伯和哉，加藤正治，小坂哲夫

２．発表標題
ツイート情報を利用した言語モデルによる感情音声認識

３．学会等名
日本音響学会春季講演論文集

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

小坂研究室
https://speech-lab.yz.yamagata-u.ac.jp/

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