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　 ６ 月 ２０ 日現在

機関番号：８２６２６
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２０５９

研究課題名（和文）異種言語感情音声コーパスの統合による多言語感情認識システムの開発

研究課題名（英文）Development of multi-lingual speech-based emotion recognition system by using heterogeneous emotional speech corpus

研究代表者
李　時旭（LEE, SHI-WOOK）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・主任研究員

研究者番号：５０４１５６４２

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）：本研究では、異種言語における、特徴正規化とマルチタスク学習に基づくシステム構築により、日本語と英語の異種言語間でも共通的な音声感情の特徴空間を構築できた点が挙げられる。特に、英語音声のみで構築されたシステムへ日本語の音声を入力する言語非依存のタスクにおいて、トリプレットネットワークにより45.05%から80.66%への35.61%の性能向上が得られた。また、ドメイン敵対的ニューラルネットワークに基づくアンサンブル手法を提案し、個別システムでは敵対的ネットワークの認識性能が、ドメイン依存性のマルチタスク学習より低い性能を示すが、提案手法による性能は逆転的に高くなることであった。

研究成果の学術的意義や社会的意義
実用化の成功が著しい音声認識分野のコーパスとは対照的に、感情音声は低資源問題とも言えるほど学習データが少ないため、実用化が未だに難解な問題であった。本研究は、多言語の感情音声コーパスから感情音声の普遍的特徴空間を構築することであり、感性コミュニケーションを実現するための核心的な研究課題として学術的な意義を持つ。また、言語、性別と感情の3つのタスクを同時に最適化するマルチタスク学習、 アンサンブル手法により、日本語と英語の両方の性能において単一システムの性能を超える多言語システムの性能が得られた研究成果は人間と共感するコミュニケーション機械の開発における社会的な意義が高いと言える。

研究成果の概要（英文）：In this study, we were able to make a common speech emotion feature space between heterogeneous languages, Japanese and English, by constructing a system based on feature normalization and multi-task learning. Particularly in language-independent tasks of inputting Japanese speech into a system built entirely of English speech, the proposed triplet network provided a 35.61% performance improvement from 45.05% to 80.66%. We also proposed an ensemble method based on a domain adversarial neural network. For the individual system, the recognition performance of domain adversarial neural networks is lower than that of domain-dependent multi-task learning, but the performance of the proposed method using an ensemble method is reversibly higher.

研究分野：知覚情報処理

キーワード：音声感情認識　音声信号処理　機械学習　パターン認識　深層学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

コンピュータの誕生とともに始まった音声認識技術は、その長期間に渡る膨大な研究成果の末、実用化を果たしてきた。最近では、Apple Siri, OK Google, Microsoft Cortana や Amazon Alexa などの対話型人工知能アシスタントが急激に普及されている。学術用語としても、人間と機械のインターフェス(human-computer interface; HCI)から人間と機械のコミュニケーション(human-machine communication; HMC)を言及することが自然になった。人間と機械の感性的なコミュニケーションを豊かにするためには、音声に基づく感情認識は不可欠な要素技術であり、最近、その関心が高まっている。人間の感情を認識し、感情と感情的な行動を人間のように合成することができる感性的知性を持つ機械を求めることはコンピュータの発展と共に 1970 年代から始まった。コンピュータの性能高度化と共に、1996 年度の国際会議 ICSLP1996 での Dellaert らの“Recognizing emotion in speech”が具体的な学術研究の初めての発表である。また、その一年前の 1995 年から米国 MIT 大学の Picard らが提唱した感性コンピューティング（affective computing）の概念もコンピュータサイエンス分野から広まり, 人間と共感するコミュニケーションの研究が数多く行われてきた。しかし、コンピュータが普及される以前の時代では、感情の音響的表現における初期研究が心理学から発展した経緯があるため、感情の定義と分類はほとんど経験的・主観的に行われた。これらの初期時代の心理学的なアプローチにより、人間の感情は各々の言語と学術的興味に応じて異質的に定義されてきた。必然的に、感情音声コーパスの開発もこのような異質的な分類定義に基づいて行われ、現在では音声に基づく感情認識の研究開発を阻害する要因の一つと考えられる。また、これらの異質的な感情音声コーパス間の統合に関する研究が大きな関心を受けた様々な要因の中、深層学習（deep learning）の導入が最も大きな要因であることは否定できない事実である。統計的推論に基づく Gaussian mixture model (GMM) – hidden Markov model (HMM)のフレームワークから, ビックデータに基づく非線形識別モデルへ遷移する大きな変化がここ 10 年の間にパターン認識と機械学習分野で現れた。 2011 年から始まった、深層学習の非線形モデルを音声に基づく感情認識へ導入したことが、話者・言語・社会・文化間の大きな変動により統計的処理を難しくして来た音声に基づく感情認識分野でのコーパス間の統合に関する研究を大きく発展させたと考えられる。特に、実用化の成功が著しい音声認識分野のコーパスとは対照的に、感情音声は低資源問題とも言えるほど学習データが少ないため、実用化が未だに難解な問題である。この低資源問題は単なる量的な問題を指すことではなく、教師あり学習に必要不可欠であるラベル情報の付与が人間により大きく変動される点がもっと大きな問題と考えられる。したがって、多言語の感情音声コーパスから感情音声の普遍的特徴空間を定義する研究は、感性コミュニケーションを実現するための核心的な研究課題の一つである。

２．研究の目的

本研究では、音声信号から言語的な意味と意図・意思・感情なとのパラ言語・非言語情報を統合できる音声に基づく感情認識技術の学術的な基盤研究を目的とする。人間は音声による感情を主観的に表現・収容する。更に、感情は文化、社会、言語などの影響を大きく受けるため、現在までに開発された様々な言語の感情音声コーパスが異なる分類のカテゴリー

を持っており、大規模な学習データを必要とする認識・分類タスクにおいては致命的な弱点となってきた。その一方、感情は言語の壁がないユニバーサル言語ともみなされる。普遍的な言語として知られている感情を機械により認識可能とすることは、音韻の言語情報に加え、パラ言語情報を本格的に実用化する音声信号に対する本質的なアプローチとも言える。文化面や言語面で非常に高い異種性を持つ日本語と英語の感情音声を対象として普遍的特徴を探求し汎用モデルを構築する試みが本研究の学術的独自性である。音声による感情認識のほとんどは、欧米の言語を対象として行われて来た。日本語を対象とする音声信号処理の歴史は長く、数多くのコーパスが存在するにもかかわらず、高い異種性のため多言語システムの開発に関する研究では日本語を含む事例が稀である。感情音声の主観的なラベルや異なる感情分類の定義などの問題を解決するため、単一言語システムの感情認識性能を超える多言語システムの構築を図る発想転換のアプローチが本研究の学術的創造姓である。多言語システムの構築が可能であると考える根拠は、先に述べたように感情がユニバーサル言語と見なされているためである。 一般的には、単一の言語のシステムの性能が優れていると知られている。しかし、過学習を防止する効果的な正規化を施すことにより、多言語システムの汎化性能を向上できると考えられる。最近の深層学習に広く適用される正規化手法としてはマルチタスク学習手法、dropout, batch normalization, データ正規化等が汎化性能の向上のため用いられる手法として挙げられる。その中、多言語の複数コーパスによるシステムの普遍性を高める多様なデータによる正規化手法が考慮できる。本研究の目的は、多言語の感情音声コーパスから普遍的な特徴を抽出する手法により、汎化問題の本質的な解決を図る研究提案である。

３．研究の方法

まず、本研究の具体的な目標として、多言語音声感情認識の性能が単一言語音声感情認識システムの性能を超えることを設定した。本研究では、英語と日本語の感情音声コーパスに加え、ドイツ語・フランス語・オランダ語などの複数の公開感情音声コーパスを整備し、共通的に使用することを計画した。個別のコーパスで定義された分類から共通するカテゴリーと個別にするカテゴリーを設定し、これらを混合する特徴空間とネットワークの最適化を初年度の研究目標とした。具体的には、一つの発声を処理単位として扱う静的特徴フレームワークと音声認識と同じく短時間の分析フレームに基づく動的特徴フレームワークの比較実験を初期段階の研究課題として行った。話者の感情や意図を把握するためには、声の高さや大きさなどの韻律情報に基づくパラ言語情報を利用することが必須である。 その中、OpenSMILE(Open Speech & Music Interpretation by Large Space Extraction)が,音声認識、音楽情報、パラ言語処理の研究向けに公開され,音声特徴と韻律特徴の統計的なパラメータが多く定義され、パラ言語情報の統計モデルを構築する研究の手掛りが提供された。本研究では、音声信号処理の国際会議である INTERSPEECH の一連のパラ言語チャレンジタスクから開発・提供された IS09 の 384 次元、IS10 の 1582 次元などをはじめとする様々なパラ言語特徴を比較実験から最適化することを研究の初期段階に行った。音声に基づく感情認識は、大きく二つに分類されるタスクを持つ。感情を分類問題として扱うカテゴリータスクと連続的な次元として扱う次元タスクである。本研究では、複数の多言語感情音声を利用するため、音声に基づく感情認識をカテゴリータスクとして設定することを初期設定とした。研究の二年目以降では、音声信号処理の最大の特徴である時系列情報を積極的に導入するため、畳み込みニューラルネットワーク（Convolutional neural networks；CNN）と回帰ニューラルネ

ットワーク (long short term memory recurrent neural networks ; LSTM-RNN)を動的特徴フレームワークに適用する。次に、低資源の感情音声コーパス間の普遍的能力を高めるためにニューラルネットワークの過学習を除去する正規化手法の実証実験を行う。転移学習に効果的な手法であるマルチタスク学習法、dropout と triplet network を利用した contrastive learning 手法、普遍的な特徴表現を構成するためのデータ正規化手法などを実験した。多言語の感情音声からは感情分類と言語識別の二つのタスクを同時に行うことが論理的に可能であるため、言語識別から敵対的学習サンプルを生成・入力し、モデルの頑健性を高める研究を行った。
最後に、本研究では音声に基づく感情認識システムの性能を、多言語の感情音声コーパスを統合的に利用出来る正規化手法として、敵対的学習（Domain Adversarial Neural Networks）のアンサンブル手法を実証し、多言語音声感情認識の性能が単一言語音声感情認識システムの性能を超える実証結果を研究成果として発表できた。

本研究では、異種言語の感情音声コーパスを統合した多言語システムの構築を以下の項目に基づいて推進した。
１）言語（コーパス）の種類を増加して、汎用性を高める手法を研究する。研究初年度の日本語（JTES; Japanese Twitter-based Emotional Speech）、英語(IEMOCAP; Interactive Emotional Dyadic Motion Capture)の二つの言語に追加して、日本語(OGVC; Online gaming voice chat corpus with emotional label)、ドイツ語(FAU-Aibo)やフランス語(RECOLA; Remote Collaborative and Affective Interactions)などのコーパスを用いて、多言語システムの汎用性を高める手法を研究した。
２）言語（コーパス）間のドメインシフトによる認識精度の劣化を究明し、感情認識タスクに共通の特徴空間を構築する手法を研究した。言語依存性のない特徴空間を多言語間で共通する特徴から構築し、言語非依存による性能劣化を防ぐことを調査した。
３）感情認識における音声特徴の静的と動的フレームワークの統合する手法を研究した。一つの発声を処理単位として扱う静的特徴フレームワーク（feed-forward neural network）と短時間の分析フレームに基づく動的特徴フレームワーク(RNN; Recurrent Neural Network) を統合する手法を研究した。その際、まずは注意機構（Attention mechanism）に基づく Sequence-to-sequence 技術に適用し、発声全体に基づくパラ言語特徴とフレームに基づくスペクトラム特徴を統合する手法を実証調査した。また、タスクである感情とドメイン情報である言語・性別などを DANN(Domain Adversarial Neural Network)と MTL(Multi-task learning)などの手法により分離または統合し、感情とドメイン情報の各々の個別特徴空間を構築する。その結果から、未知のドメインへ頑健な汎用特徴空間（表現）を構築するためのデータ拡張(data augmentation)を最適に適用する手法の研究・開発を行った。

４．研究成果

研究初年度として、複数の公開感情音声コーパスを整備し、共通的に使用することを進めた。日本語は JTES（Japanese Twitter-based Emotional Speech）を、英語は音声感情認識の学術分野で共通のベンチマークテストとして広く用いられる IEMOCAP（Interactive Emotional Dyadic Motion Capture)を用いた。まず、個別のコーパスで定義された分類から共通するカテゴリーを設定し、これらを混合する特徴空間、特徴正規化(feature normalization)とマルチタスク学習（Multi-task learning; MTL）に基づくネットワークの最適化を初年度の研究課題

として進めた。その結果、異種性の高い日本語と英語の感情音声データを用い、音声特徴のみに基づく感情認識の高性能深層ニューラルネットワークを構築できた。多言語による汎化性向上の研究として混合特徴空間とネットワークの最適化の実証実験を進めた。その研究成果として、論文投稿時点では世界最高性能の音声感情認識の正解率が得られ、google scholar の Acoustics & Sound のトップ 1 位の国際会議である ICASSP2019 で発表した。

研究二年目では、新たに英語の感情音声データベース(MSP-IMPROV)を加えて、三つの多言語のクロス言語間の認識タスクを設定し、言語間における共通的な感情特徴空間を構築することを進めた。三つの異種感情音声データを対象にして、近年注目の高い triplet network を用い、感情における共通・汎用空間を探索した。研究の結果、二つの英語音声データから構築されたモデルに対し、日本語の感情音声を入力した際、言語間の相違により 45.05%の低い正解率であった性能を、提案の triplet network を用いる手法では、言語間の学習データの併用なしでも 80.66%まで 35.61%の性能向上を果たすことが出来た。この結果から、異種言語間でも共通する汎化性の高い特徴空間が存在することが確認できた。この研究成果は、google scholar の Acoustics & Sound 分野におけるトップクラスの国際会議である IEEE Spoken Language Technology Workshop(SLT2021)で採択され、発表を行った。
この研究進捗は、研究初年度の異種言語における、特徴正規化(feature normalization)とマルチタスク学習（Multi-task learning; MTL）に基づくシステム構築に続き、三つの多言語のクロス言語間の認識タスクを設定し、異種言語間でも共通的な音声感情の特徴空間を構築できることを確認した点であった。特に、英語音声データのみで構築されたシステムへ日本語の音声を入力する言語非依存（独立）の実証実験において、提案の triplet network により 45.05%から 80.66%への 35.61%の性能向上を得られたことは大きな研究進捗であった。

研究最終年度の三年目では、日本語感情音声コーパスと英語感情音声コーパスを対象とする多言語音声感情認識において、ドメイン敵対的ニューラルネットワーク(domain adversarial neural network; DANN)をアンサンブルする手法を提案し、性能向上の成果が得られた。これは、個別システムではドメインへの依存性を低く抑えた DANN の認識性能が、補助タスクのない普通システム及びドメインへの依存性を強化したマルチタスク学習(multi-task learning; MTL)の性能より低い性能を示すが、複数システムを融合するアンサンブルによる性能は逆転的に高くなることである。即ち、特定タスクである感情以外の言語や性別などの情報を除去する DANN によってタスクに関連する情報も毀損されたが、アンサンブルによって複数言語に渡る共通因子が抽出でき、汎化と識別の両方の性能を兼ね備えた特徴空間が構築できたと考えられる。この研究成果は、google scholar の Acoustics & Sound 分野におけるトップクラスの国際会議である IEEE Automatic Speech Recognition and Understanding Workshop (ASRU2021)で採択され、発表を行った。

三年間の研究開発を通して、本研究では、言語、性別と感情の 3 つのタスクを同時に最適化するマルチタスク学習、 contrastive learning と triplet network のアンサンブル手法により、日本語と英語の両方の性能において単一システムの性能を超える多言語システムの性能が得られた。その成果をトップの国際会議で毎年発表することができたため、本研究課題の進捗はおおむね順調であった。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2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 伊藤 慶明、岩崎 瑛太郎、金子 大祐、小嶋 和徳、李 時旭	４．巻 J103-D
２．論文標題 音声中の検索語検出における音声クエリ・音声ドキュメントのフレームレベル最ゆう系列化照合方式	５．発行年 2020年
３．雑誌名 電子情報通信学会論文誌D 情報・システム	６．最初と最後の頁 919～928
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2020JDP7030	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 －

１．著者名 丹治遥，小嶋和徳，李時旭，南條浩輝，伊藤慶明	４．巻 61
２．論文標題 音声中の検索語検出におけるクエリの関連語を利用したリスコアリング方式	５．発行年 2020年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 103~112
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7件（うち招待講演　0件／うち国際学会　6件）

１．発表者名 Shi-wook Lee
２．発表標題 DOMAIN GENERALIZATION WITH TRIPLET NETWORK FOR CROSS-CORPUS SPEECH EMOTION RECOGNITION
３．学会等名 2021 IEEE Spoken Language Technology Workshop (SLT)（国際学会）
４．発表年 2021年

１．発表者名 Takashi Yokota, Kazunori Kojima, Shi-wook Lee, Yoshiaki Itoh
２．発表標題 Reduction of Speech Data Posteriorgrams by Compressing Maximum-likelihood State Sequences in Query by Example
３．学会等名 APSIPA-ASC2020（国際学会）
４．発表年 2020年

１．発表者名
西野将弘，小嶋和徳，李時旭，伊藤慶明

２．発表標題
異種・複数の深層学習モデルを用いた音声中の検索語検出方式の高精度・低メモリ化

３．学会等名
日本音響学会春季研究発表会

４．発表年
2021年

１．発表者名
H. Tanji, K. Kojima, H. Nanjo, S. Lee, and Y. Itoh

２．発表標題
A Rescoring Method Using Web Search and Word Vectors for Spoken Term Detection,

３．学会等名
APSIPA-ASC2019（国際学会）

４．発表年
2019年

１．発表者名
Shi-wook Lee

２．発表標題
THE GENERALIZATION EFFECT FOR MULTILINGUAL SPEECH EMOTION RECOGNITION ACROSS HETEROGENEOUS LANGUAGES

３．学会等名
ICASSP-2019（国際学会）

４．発表年
2019年

１．発表者名
Kazuki Hatakeyama, Masahiro Nishino, Kazunori Kojima, Shi-wook Lee, Yoshiaki Itoh

２．発表標題
Multiple Deep Learning Models and Architectures with Different Numbers of States Used to Improve Retrieval Accuracy of Query-by-Example

３．学会等名
APSIPA-ASC2021（国際学会）

４．発表年
2021年

１．発表者名 Shi-wook Lee
２．発表標題 ENSEMBLE OF DOMAIN ADVERSARIAL NEURAL NETWORKS FOR SPEECH EMOTION RECOGNITION
３．学会等名 IEEE AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING WORKSHOP (ASRU2021)（国際学会）
４．発表年 2021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