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　６ 年　 ５ 月 ２４ 日現在

機関番号：３２４０７
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１９４１

研究課題名（和文）複数センサの融合による発声動作からの発話内容の推定と発声補助デバイスへの応用

研究課題名（英文）Estimation of speech content from vocal movements by fusion of multiple sensors and its application to speech assistance devices

研究代表者

大田　健紘（Ota, Kenko）

日本工業大学・基幹工学部・助教

研究者番号：５０５１１９１１

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：研究期間全体を通して本研究では、声帯を除去するなど発声が困難となった人の発話の補助や、既存の音声認識を補助するシステムの検討を目的とした。その結果、音声情報を用いることなく音素単位で文章を認識する深層学習を用いた技術について検討ができた。そして、発声補助方法として、カメラ及び指の皮膚電気抵抗を計測するセンサを用いた感情の推定技術やテキストからの音声合成技術について取り組み精度向上への基礎検討ができた。

研究成果の学術的意義や社会的意義
本研究は、音声情報を利用しない音声認識について、深層学習を用いた音素単位での文章認識を実現するためのデータ取得手法や深層ニューラルネットワークについて検討したことに学術的な意義がある。また、話者の感情推定や音声合成技術それぞれについて取り組み、発声が困難な方のための発声補助デバイスの開発に向けた基礎的な検討ができたことや課題の抽出ができたことに社会的な意義がある。

研究成果の概要（英文）：Throughout the entire research period, the purpose of this study was to investigate systems that assist people who have difficulty speaking, such as by removing their vocal cords, and systems that assist existing speech recognition. As a result, we were able to study a technology using deep learning that recognizes sentences phoneme by phoneme without using speech information. We also studied a technology for estimating emotions using a camera and a sensor that measures the galvanic skin response of the fingers, and a speech synthesizing technology from text as a method for assisting speech production.

研究分野：知能情報処理

キーワード：無発声音声認識　深層学習　三次元計測

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

現在，音声認識技術は深層学習の登場により，その認識率は従来の認識手法と比較して格段に向上した．その結果，利用状況は限定されるが様々な音声認識の応用システムが実用化され，我々の生活の中で活用されている．音声によるヒューマン・マシンインターフェースは成功しつつあるが，音声を発する必要があることから，

・声帯を摘出し音声を発することができない人には利用することができない，
・雑音や残響が酷い環境では未だに利用は困難，
・プライバシーを保護することができない

などの問題点がある．そのため，新しいヒューマン・マシンインターフェースとして，音声情報を利用しない手段が必要とされている．

２．研究の目的

我々は，高雑音環境下における音声認識の補助技術という観点だけではなく，喉頭癌などの病気に罹患し発声器官を除去したことにより発声が困難となった方のコミュニケーションの補助技術として音声情報を利用しない無発声音声認識の研究に取り組む．

（１）無発声音声認識を実現するために，我々は口の周囲の筋電位や熱画像，口の動画像を用いて単語を認識する手法を検討し，それらの有効性を明らかにする．

（２）また，口の動画像を用いて日本語の文章の認識手法について検討し，認識性能を明らかにする．

（３）さらには，発声補助デバイス開発の基礎検討として，話者の感情推定及びテキストからの音声合成についても取り組み，デバイス開発に向けた課題を明らかにする．

３．研究の方法

（１）まず，音声情報を利用せず口の動作をもとにした無発声音声認識を実現する手段としては，カメラ（可視光、熱画像）を利用する方法や筋電位を利用する方法などがある．これらにより取得したデータを入力とした深層学習に基づく無発声での単語認識を行う．これにより認識性能を評価し，無発声での単語認識に有効な入力データについて検討する．

（２）カメラにより取得したデータを用いた無発声での文章認識を実現するためには，非常に多くのデータと発話内容を書き起こしたテキストデータが必要となる．しかしながら，データの収集は非常に手間のかかる作業である．そのため，カメラで取得したデータの水増しを行う手法を導入し，深層学習による音素単位での文章認識を行う。これにより，学習中の損失の推移を確認し，認識性能を評価することで，文章認識に適した深層ニューラルネットワークについて検討する．

（３）話者の感情推定については音声を用いることができないため，生体情報に基づく方法を採用する．カメラで撮影した映像の色の変化を解析することで心拍情報を得ることができる映像脈波抽出技術を採用する．これに加えて，皮膚の電気抵抗の変化を計測するセンサも導入する．音声合成については，合成対象の話者の音声を大量に収集できない場合であっても，高い品質で合成可能とするために，転移学習に基づく手法を検討する．

４．研究成果

（１）可視光カメラ及びサーモグラフィー，筋電位センサを用いて，発話動作から口の動きや口内の情報に関する時系列データを取得し，それらを融合する深層学習を行った．その結果，サーモグラフィーにより得られた熱動画像を学習データに含むことが認識性能の向上に寄与する傾向がみられた．ただし，学習及び評価に用いたデータ数が不十分であるため、引き続きの検証が必要である[1].

（２）以上の研究と並行して可視光カメラにより得られたデータを用いた無発声での単語認識についても検討した．十分なデータを収集できない場合，深層学習により認識モデルの学習を十分に行えない．さらには，従来の無発声音声認識では正面を向いた状態での発話を想定しており，より現実的な状況を想定した場合，顔の向きや体の動きにより十分な認識性能を得られなくなる可能性がある．そのため，カメラで撮影した顔画像をもとに顔の３次元モデルを作成し，それをさまざまな方向へ回転させることで学習データを増加させるとともに，正面以外の方向を向いて発話したとしても認識可能となる手法を検討した．まずは単眼のカメラで３次元モデルを作成可能な DECA（Detailed Expression Capture and Animation）を用いて学習データを生成した．畳み込みニューラルネットワークを有する深層学習モデルにより，日本語の日常会話で用いられる無言も含めた 11 単語について単語識別を行ったところ約 80%の精度が得られた[2].

しかしながら，DECA を用いた方法はデータ生成に時間がかかるため，文章を認識するため

に大量の学習データを準備することには不向きであった．そこで，より短時間で３次元モデルを作成するために，カメラを２台使うことになるが，３次元計測法の一つである DLT（Direct Linear Transformation）法を用いた．図 1 に DLT 法により作成した唇の３次元モデルを示す．唇の特徴点 20 点分の３次元座標の時系列データを学習データとして用いた．このデータを入力として，DeepSpeech2 をベースとする深層ニューラルネットワークを用いて音素単位での文章認識を行った．その結果，ネットワークの学習に用いた話者が学習に用いた文章を評価のために録画したデータではあるが，検討した条件の中で最も低い音素誤り率を得ることができた．しかしながら，学習に用いた文章とは異なる場合は，大幅に性能が低下することもわかった[3]．図２に比較的正確に認識できた例を示す．なお，テキストコーパスとして ITA コーパスを利用した[4].

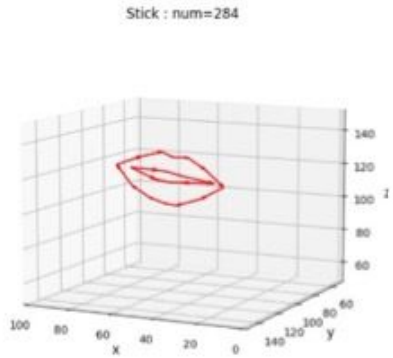

図１　３次元モデルの一例

```
REF: sil s u p i r i cl ts u t o w a j o o r y u u sh u n    o k o t o d e s u sil
HYP: sil s u p i r i cl ts u t o w a j  o r y u u sh u n o k o k o t o d e s u sil
EVA:                                D              I I
PER: 9.09%

REF: sil s u p i r i cl ts u t o w a j o o r y u u sh u n o k o t o d e s u sil
HYP: sil s u p i r i cl ts u t o w a j o o r y  u sh u n o k o t o d e s u sil
EVA:                                      D
PER: 3.03%

REF: sil s u p i r i cl ts u t o w a j o o r y u u sh u n o k o t o d e s u sil
HYP: sil s u p i r i cl ts u t o w a j o o r y  u sh u n o k o t o d e s u sil
EVA:                                      D
PER: 3.03%
```

朗読文：スピリッツとは蒸留酒のことです
REF：目標値、HYP：予測値、EVA：判定、S：置換、I：挿入、D：削除

図２　認識結果の例

（３）発声補助デバイス開発には，無発声音声認識技術だけではなく，話者の感情推定や音声合成技術の検討が必要となる．まず，感情推定は心拍や皮膚電気抵抗などの生体情報を用いて行われる．デバイスの利用者に負担にならないことを考慮し，カメラで撮影した顔画像の色の変化をもとに心拍変動を抽出する映像脈波抽出技術及び指先から皮膚電気抵抗を測定する方法を検討した．しかしながら，心拍変動の推定精度が悪いこともあり，十分な推定精度を得ることはできなかった．そのため，照明環境の変化に頑健な動的モード分解を用いた脈波抽出技術を導入するなど，心拍変動の推定精度向上が課題となる．次に，音声合成に関しては，テキストからメルスペクトログラムを生成する tacotron2 及び，メルスペクトログラムから音声波形を生成する waveglow を組み合わせた方法を用いた．合成対象話者が日本語の文章（300 文程度）を発話した音声データを用いて転移学習を行ったところ，比較的短い言葉であれば収録した合成対象話者の音声に近い主観評価結果となった．そのため，今後は合成対象話者のデータ数をさらに少なくし、さまざまな文章を発声可能とする学習方法の導入が課題となる．

参考文献

[1] 草本 雅也, 大田 健紘, "複数のセンサを用いる無発声単語認識に関する研究", 信学技報, vol. 121, no. 404, MICT2021-104, pp. 19-24, 2022 年 3 月.

[2] 和田 竜二, 大田 健紘, "深層学習に顔の 3 次元モデルを用いた無発声単語認識に関する研究", 信学技報, vol. 121, no. 404, MICT2021-103, pp. 13-18, 2022 年 3 月.

[3] 大田 健紘, 久保　茜, 倉島　廉, "口唇特徴点の時系列データに基づいた日本語機械読唇手法の検討", 信学技報, vol. 123, no. 355, MICT2023-48, pp. 52-57, 2024 年 1 月.

[4] 小口 純矢, 金井 郁也, 小田 恭央, 齊藤 剛史, 森勢 将雅：ITA コーパス：パブリックドメインの音素バランス文からなる日本語テキストコーパスの構築と基礎評価, 情報処理学会研究報告, vol. 2021-MUS-131, no. 31, pp. 1-6, 2021.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 大田健紘、久保　茜、倉島　廉
２．発表標題 口唇特徴点の時系列データに基づいた日本語機械読唇手法の検討
３．学会等名 電子情報通信学会ヘルスケア・医療情報通信技術研究会
４．発表年 2024年

１．発表者名 Kenko Ota
２．発表標題 Silent speech recognition using data augmentation based on a 3D lip model
３．学会等名 Acoustical society of America（国際学会）
４．発表年 2023年

１．発表者名 木村一馬，大田健紘
２．発表標題 機械読唇における三次元モデルを用いたデータ拡張が認識精度に与える影響
３．学会等名 電子情報通信学会ヘルスケア・医療情報通信技術研究会
４．発表年 2023年

１．発表者名 和田竜二，大田健紘
２．発表標題 深層学習に顔の3次元モデルを用いた無発声単語認識に関する研究
３．学会等名 電子情報通信学会ヘルスケア・医療情報通信技術研究会
４．発表年 2022年

１．発表者名 草本雅也，大田健紘
２．発表標題 複数のセンサを用いる無発声単語認識に関する研究
３．学会等名 電子情報通信学会ヘルスケア・医療情報通信技術研究会
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