研 究 終 了 報 告 書

「口腔内超音波画像と深層学習を用いた無声発話認識に関する研究」

研究期間： 2019 年 10 月～2022 年 3 月
研 究 者： 木村　直紀

1. 研究のねらい

音声によって操作するインタフェースは基本的に誰でも簡単に使える次世代のインタフェースです。しかし発声を必要とするため利用可能な場面やユーザが限られてしまう問題があります。そこで本研究では、発声することなく調音器官を動かし、その時の調音運動を計測・解析することで発話内容を推定する、「サイレントスピーチ」による解決を目指します。またサイレントスピーチを用いたコンピュータ・インタラクションを提案します。

2. 研究成果

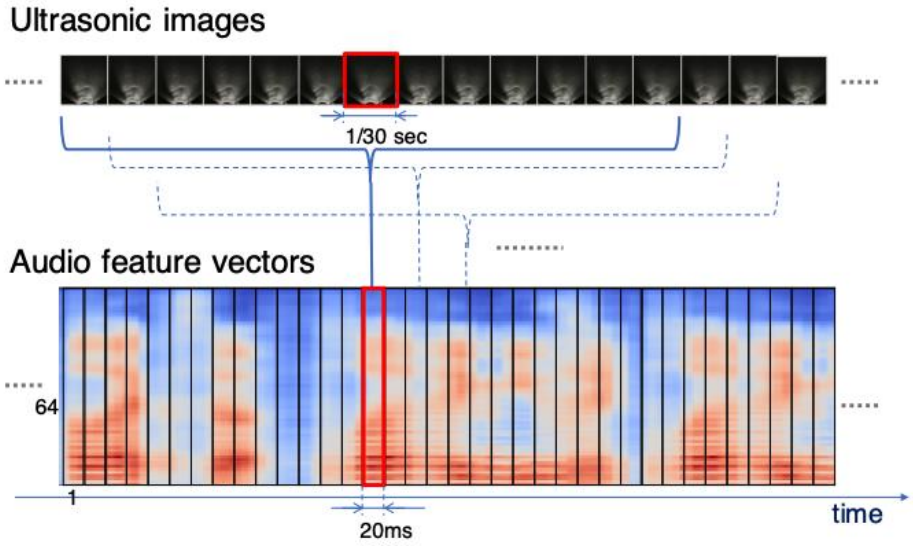
（1）概要

本研究では採択前に構築された深層学習フレームワークの性能向上に取り組み，軽量化と語彙数の上昇に成功した（研究テーマ A）．これを読唇用に応用したウェアラブルサイレントスピーチインタフェースを開発し，国際学会 AVI2020 で採択・発表された．さらに大きな語彙に対応可能にするために，音響特徴量を直接推定するアプローチから，音声認識のフレームワークを使うアプローチにピボットした．音声認識のコミュニティで近年性能を発揮している，connectionist temporal classification(CTC)と attention based encoder-decoder network という機構を用いて，Silent Speech Challenge（SSC）と呼ばれる超音波画像と唇画像がセットになったデータセットのベンチマークにチャレンジした．この結果は，国際学会 INTERSPEECH2020 にて発表された．この試みから，より大きなデータセットを集めることで，End-to-End モデルの性能をより引き出すことができると示唆されたため，SSC の約3倍のサンプルサイズを持つデータセットを新たに構築した．

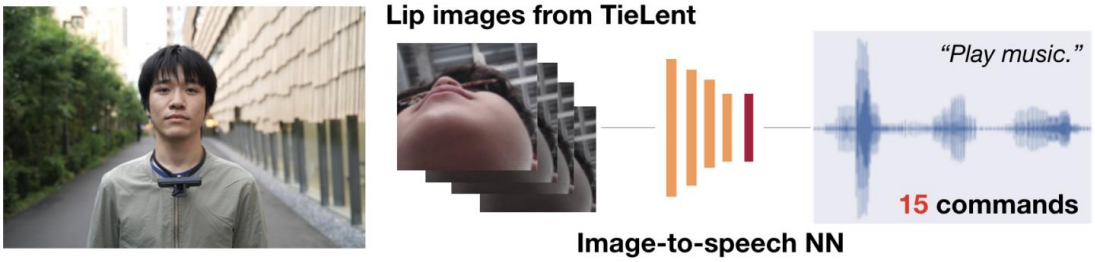
（2）詳細

研究テーマ A「超音波エコー画像と深層学習による無発声ボコーダ」

本研究は，採択される前に行っていた事前実験とその成果に基づいて提案されたものであった（https://dl.acm.org/doi/10.1145/3290605.3300376）．この研究では，ある発音とその時の口の動き数フレームが対応しているという野心的な仮定を置き，実際に 10 フレーズほどの発話が可能であることを２人の実験参加者で確かめたものだった．この仮定に基づいて構築されていたニューラルネットワークは，原理的には，語彙の制限のない open vocabulary を実現し得るものであり，ニューラルネットワークの技術的な熟練によって解決できることを目指していた．しかし，ネットワークの改良とデータセット収集によって 30 フレーズ程度の語彙までは到達したものの，それ以上の語彙に増やすと識別不可能な音声を出力してしまうようになった．



研究テーマ B「唇画像と深層学習による無発声ボコーダを用いた，ウェアラブルサイレントスピーチインタフェース」

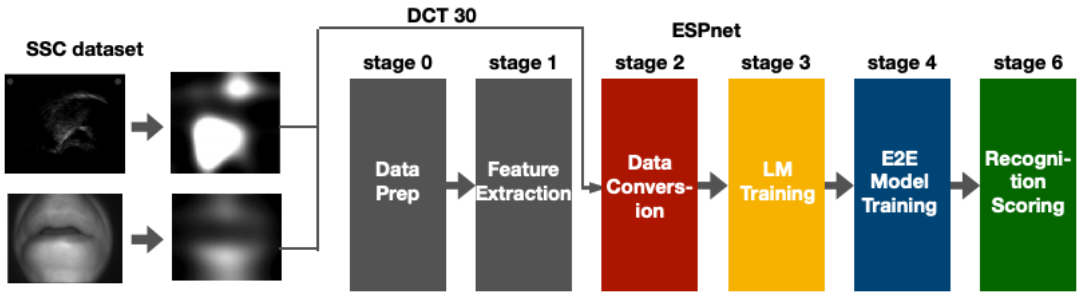


研究テーマ A は当初の目論見に失敗したものの，オフライン条件下で 30 フレーズ程度までなら明瞭な（人間が識別可能な）音声を生成することができた．この技術を唇画像に対して様々な実験をしている際に発見した，「唇が写っている限り，生成音声の質はほとんど変わらない」という発見を用いて，安価な web カメラを首元に装着可能し，音声を生成するサイレントスピーチインタフェースを提案した．ニューラルネットワークの部分は研究テーマ A で開発されたものであるため，音声を出力するが，論文では classification タスクによっても評価し，15 コマンドの語彙のなかで平均 94%の accuracy を記録した．この研究は，正面以外からの lip reading（もしくは lip image to audio）という新しいタスクを提案している点が評価され，国際学会 AVI2020 にてフルペーパーとして採択された（主な研究リスト 1）．

研究テーマ C 「End-to-End Deep Learning Speech Recognition Model for Silent Speech Challenge」

研究テーマ A の結果を踏まえ，語彙数の増加を目指し，深層学習を用いるという点には軸足を残しつつ，音声認識の考え方や技術・パイプラインを用いる方針を採用した．超音波画像をベースとしたサイレントスピーチ認識においては，Silent Speech Challenge（以下 SSC）というデータセットが存在している．SSC には 2342 回の TIMIT をベースとした発話が training data, 100 回

の WSJ0 をベースとした発話が test data として用意されている. これまでに Kaldi を用いた DNN-HMM がベンチマークとして記録されていたが, 音声認識では connectionist temporal classification(CTC)や attention based encoder-decoder network が state-of-the-art になりつつあり, SSC に対してはまだ試されたことはなかった. 本研究では ESPnet (end-to-end speech processing toolkit)を用いて, hybrid CTC / attention based end-to-end ASR model を SSC に適用した. 画像特徴量としては Discrete Cosine Transform(DCT)とオートエンコーダ(AE)を使用した.



結果:

2342 samples は End-to-End モデルを訓練するのには小さすぎるため, ESPnet に内蔵された, SpecAugment と呼ばれるデータオーグメンテーション手法を用いた. このオーグメンテーションによる結果の改善は著しく, Character Error Rate(CER)がほぼ半分に削減された.

Table 1: *Recognition results (CER) with and without SpecAugment*

SpecAug	Sub (%)	Del (%)	Ins (%)	CER (%)
Yes	4.2	5.9	1.9	**10.1**
No	8.1	10.5	4.0	18.7

SpecAugment を使用した上で, DCT, AE 二つの特徴量と次元数による結果の比較を行った. より新しい手法である AE による改善を期待していたが, 最終的には古典的な特徴抽出手法の優位性が示される結果となった.

Table 2: *Recognition results (CER and WER) of different features with SpecAugment*

Feature	Dimension	CER (%)	WER (%)
DCT	20	17.1	33.8
	30	**10.1**	**20.5**
AE	20	19.9	36.3
	30	17.0	33.2
	60	21.4	42.0

議論:

本研究で用いた SSC データセットの SOTA は DNN-HMM を用いた場合の 6.45%であり, 本研究の結果はそれに届いていない. ただし以前の結果は 5000words で訓練された言語モデルを用

いている．一方で，私たちの結果は 65000words で訓練された言語モデルを用いたものであり，少し不利な条件である．また，End-to-End 手法の特性（大きなデータセットを必要とする）と，SpecAugment で大きな改善が見られたことから，生データのサンプルサイズを大きくするは更なる改善をもたらすと予想された．この結果は，INTERSPEECH2020 で発表された（主な研究リスト 2）．

研究テーマ D「A Synchronized Corpus of Ultrasound Tongue Imaging for End-to-End Silent Speech Recognition」

研究テーマ C での結果を元に，SSC よりもさらに大きいコーパスを整備することは，本研究課題にとって重要であると同時に，サイレントスピーチコミュニティ全体にとって重要だと考えた．近年 Lip Reading(読唇)のタスクには音声認識や画像認識をバックグラウンドにもつ研究者が多く参入している．これは Grid Corpus や Lip Reading in the Wild など，大規模コーパスが整備されたことが重要であった．

私たちは一人のネイティブスピーカーを募集し，合計で 7384samples を収集した．これは SSC の約3倍であり，一人のスピーカーに対しての唇画像データセットとしても過去最大のものである．

3. 今後の展開

・サイレントスピーチ認識パイプラインの公開

本研究課題で開発されたサイレントスピーチ認識パイプラインは非常に強力なパッケージであり，あらゆるモダリティ（入力データの次元数が自由）に対応できる．ヒューマン・コンピュータ・インタラクション（HCI）の分野では，近年サイレントスピーチインタフェースに関する研究が急増しているが，認識手法は共有されておらず，各々で作成されている．サイレントスピーチ認識に関しては，2010 年頃から音声認識のコミュニティで開発されてきているが，コミュニティが断絶しており，その知見が共有されることもない．HCI コミュニティには，技術開発だけでなく，ユーザ・エクスペリエンスや social acceptability などの議論が期待されるが，技術開発から始める必要があるために，未だそれらについて議論した論文は少ない．本研究のパイプラインを HCI コミュニティで共有することで，HCI におけるサイレントスピーチの議論を深めることができると考えている．

・収集したデータセット公開

収集した新たなデータセットを生データ，前処理後のデータ，認識パイプラインのセットとして公開する．これによって，前処理に取り組みたい人，認識器の改善に取り組みたい人，モチベーションの異なる人が取り組みやすい環境を作る．また，最終的には読唇に注がれている投資や人材をサイレントスピーチ認識に惹きつけることが重要と考えており，読唇データセットとして代表的な Lip Reading in the Wild, Lip Reading Sentences 等のデータセットとの相互リンクを貼ることを目指す．

・Self-supervised learning for silent speech recognition and human-computer Interaction

近年，音声認識では Self-supervised learning と呼ばれる手法が盛んに開発されている．これは少量のラベル付きデータから音声認識を可能にする技術である．この技術は既存のベンチマークでも有効であるが，特にラベル付きデータの少ない少数言語に有効だとされている．そしてこの特性はラベル付きデータを集めることが困難なサイレントスピーチにおいて非常に強力であることが予想される．さらに Human-Computer Interaction 全体として，ラベル付きデータを集めるということは課題であり，Self-supervised learning は HCI 分野の様々なタスクにおいて汎用的に有効な手段になると考えている．本研究者は，今後 3 年程度のタイムスパンを見据えて，この課題に取り組む予定である．

4. 自己評価

まず，「口腔内超音波画像と深層学習を用いた無声発話認識に関する研究」という研究課題に対して，アドバイザーの方々のお力添えのもと，正面から取り組み，ある程度の答えを出すことができたと考えている．ただし，既存のベンチマーク（Silent Speech Challenge）を明確に超えることができなかったこと，そしてその後のパイプラインとデータセットについて論文という客観的な成果を出せていないことは，研究者の至らなかった部分だと感じている．

研究の進め方は良くも悪くも，堅実なものだったと感じている．この課題に採択されてからの3年間，サイレントスピーチインタフェースという一つのテーマに研究費も研究者の時間も費やされた．その姿勢は一定の成果を出すことにつながったものの，課題以外のコンピュータサイエンスに関する知識を身につける時間がなく，また適切に休憩を取ることができず，体調を崩してしまい，結果的に生産性が落ちるという結果になった．ただし，本研究者の最初の競争的資金であり，良い学びになったと考えている．

今回開発されたサイレントスピーチ認識パイプラインは，共有し，適切に PR を行うことで，Human-Computer Interaction 分野におけるサイレントスピーチ研究のフェーズを一歩進めることができると自負している． また研究者自身は，speech のコミュニティと HCI のコミュニティの両方につながりのある数少ない研究者となったことで（この二つは分断されていることが多い），相互の交流・良いシナジーを生み出せるのではないかと考えている．

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：2件

1. Naoki Kimura, Kentaro Hayashi, Jun Rekimoto . TieLent: A Casual Neck-Mounted Mouth Capturing Device for Silent Speech Interaction, Proceedings of the International Conference on Advanced Visual Interfaces , 2020/9/28, 1-8
With the increased use of smart speakers, silent speech interaction (SSI) is attracting attention. Unfortunately, traditional silent speech interaction methods require the addition of obtrusive sensors and devices around the user's face, making wearability and portability a challenge. Considering that most uses for smart speakers do not require many words, we

suggest a more casual approach, TieLent, which can easily be worn between the neck and the chest. TieLent's RGB camera is set away from the user's face, presenting less interference with the user. Although TieLent's camera is not able to capture the whole mouth, when combined with our image-to-speech neural network model, it is able to generate the recognizable speech of 15 commands with an average accuracy of 94%.
2. Naoki Kimura, Zixiong Su, Takaaki Saeki End-to-End Deep Learning Speech Recognition Model for Silent Speech Challenge, INTERSPEECH, 2020, 1025-1026
This work is the first attempt to apply an end-to-end, deep neural network-based automatic speech recognition (ASR) pipeline to the Silent Speech Challenge dataset (SSC), which contains synchronized ultrasound images and lip images captured when a single speaker read the TIMIT corpus without uttering audible sounds. In silent speech research using SSC dataset, established methods in ASR have been utilized with some modifications to use it in visual speech recognition. In this work, we tested the SOTA method of ASR on the SSC dataset using the End-to-End Speech Processing Toolkit, ESPnet. The experimental results show that this end-to-end method achieved a character error rate (CER) of 10.1% and a WER of 20.5% by incorporating SpecAugment, demonstrating the possibility to further improve the performance with additional data collection.

（2）特許出願

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

一般財団法人デジタルコンテンツ協会 Innovative Technologies 2019「深層学習によるサイレントスピーチインタラクション」