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　 ２ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2019～2021
課題番号：１９Ｈ０４１４８

研究課題名（和文）深層学習による無音声発話インタラクションの研究

研究課題名（英文）Research on Silent interaction with deep neural networks

研究代表者
暦本　純一（Rekimoto, Jun）

東京大学・大学院情報学環・学際情報学府・教授

研究者番号：２０４６３８９６

交付決定額（研究期間全体）：（直接経費）　　13,400,000円

研究成果の概要（和文）：音声インタフェースは急速に普及してきているが、公共環境や騒音環境で利用できないなどの制限があった。本課題では、深層学習による無音声発話認識の研究を行った。顎の下側に取り付けられた超音波イメージングプローブによって観察される口腔内映像から発声内容を認識する深層学習器、皮膚運動を顎や喉に添付した加速度センサーから発話を推定する機構、マスクに添付した加速度センサーにより発話を認識する機構を構築し、スマートスピーカーなど音声対話システムを駆動できることを確認した。さらに、視線情報と口唇映像からのコマンド認識を組み合わせるマルチモーダルインタフェースの構築に成功した。

研究成果の学術的意義や社会的意義
本研究成果により、音声インタラクションが公共環境や騒音環境で利用できないなどの従来の制限を超えて利用できる可能性が出てきた。音声インタラクションは他の入力手段と比較しても高速で、手指を拘束しないなどの特徴を持つ。未来のモバイルインタフェースやウェアラブルインタフェースの手段としてサイレントスピーチが広範に利用される可能性がある。さらに、咽頭障害、声帯機能障害、高齢による発声困難者に対して、声によるコミュニケーションを取り戻すための支援技術としての利用意義がある。

研究成果の概要（英文）：Speech interfaces are rapidly becoming popular, but there are some limitations, such as their inability to be used in public or noisy environments. In this project, we studied silent speech recognition using deep learning. We constructed a deep learner that recognizes speech content from intraoral images observed by an ultrasonic imaging probe attached to the underside of the jaw, a mechanism that estimates speech from acceleration sensors attached to the jaw and throat for skin movement, and a mechanism that recognizes speech by acceleration sensors attached to a mask. We confirmed that the system can drive a smart speaker or other spoken dialogue system. Furthermore, we succeeded in constructing a multimodal interface that combines eye gaze information and command recognition from lip images.

研究分野：人間拡張学

キーワード：サイレントスピーチ　人間拡張　人工知能　音声処理　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

音声対話によって制御可能なデジタル機器が非常に多くの状況で使用されるようになっている。スマートフォン、スマートスピーカー、カーナビゲーションシステム、など様々なデジタル機器は音声で制御可能になり、音声認識技術の進歩と音声合成の自然さにより、インタラクション手段としての音声対話の重要性が増している。音声対話は視覚的な注意を必要とせず、寝室などの暗い環境でも使用できる。 利用者が運転、家事、医療、移動、または従来の PC の使用などの他の作業を行っていて手が離せない状況でも使用できる。 例えば、PC 画面およびキーボードやマウスなどの対話デバイスに集中しているときにも、音声対話を用いて他のデバイスを操作することができる。さらに、視覚障害者や手指の制御が困難な利用者にとっても音声インタラクションの意義は大きい。

しかし、公共の場所での音声の使用には制限がある。 周囲の人に迷惑なることなどから発声を懸念してしまう。また個人情報や機密情報を発声することはできない。騒音環境では音声認識の精度が落ちる場合がある。これらの問題は、ウェアラブルまたはモバイルコンピューティングとの対話手段として音声対話を使用しようとする場合に特に深刻となる。

２．研究の目的

これらの課題を解決するために、本研究では無音声発話（サイレントスピーチ）に着目する。利用者が、実際には声帯を振動させずに、発話のときと同様に口や舌を動かしたときに、その発話内容を認識できれば、発声しない音声対話が可能になる。利用者がすでに有している発音や発話の能力を転用でき、音声対話可能なスマートデバイスとの親和性も高い。骨伝導イヤフォンやオープンカナル型（外耳道を塞がない）イヤフォンと、無音声発話技術を併用すると、外部に音を漏らすことなく、デジタル情報にアクセスできる。これは新しいウェアラブル・コンピュータの構成方式となる。さらに、声帯損傷や、高齢により充分な声量で発話できない人のために、コミュニケーションを補助する技術としての価値もあり、高齢化社会に向けて非常に重要な技術といえる。

本研究では (1) 超音波画像を用いて、無音声発話から音を復元する深層学習モデルを実現し、外部の音声制御デバイス(スマートスピーカー等)と連携可能なことを示す。（２）無音声発話から復元した音声を利用者が聞くことで、発話の品質が向上することを検証する。（３）超音波映像以外の手段（口唇画像、皮膚加速度）と深層学習による認識によっても、無発声音声認識が可能であることを示す。ことを目的とする。

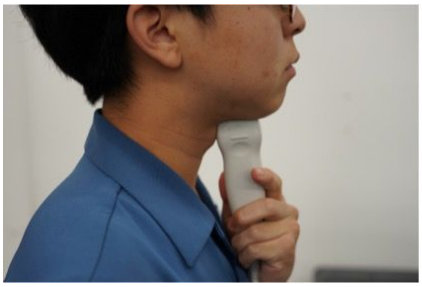
３．研究の方法

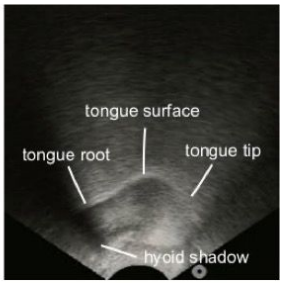
本提案では超音波イメージング・口唇画像、皮膚加速度などの無発声発話に誘発されて観測される現象を深層学習により認識し、音声復号および音声認識を行うことを研究手法とする。 超音波イメージング(超音波エコー)技術は、体内に放射される超音波の反射時間を測定することで体内の内部状態を認識するもので、医療目的で広く使用されている。 近年では、スマートフォン程度の外形で携帯可能な小型軽量のシステムも登場している(General Electronic Company Vscan Extend 等)。顎の下付近に小型超音波イメージングプローブを取り付け、口腔内の状況を計測し、それを音響情報に変換することが可能ならば、実際に声帯を振動させずに話すことで音声対応装置を制御するための有用な構成となり得る。また、復号した発話やコマンド等を外部に露出することなく、直接ウェアラブル機器等を制御する手段として利用することも可能である。

４．研究成果

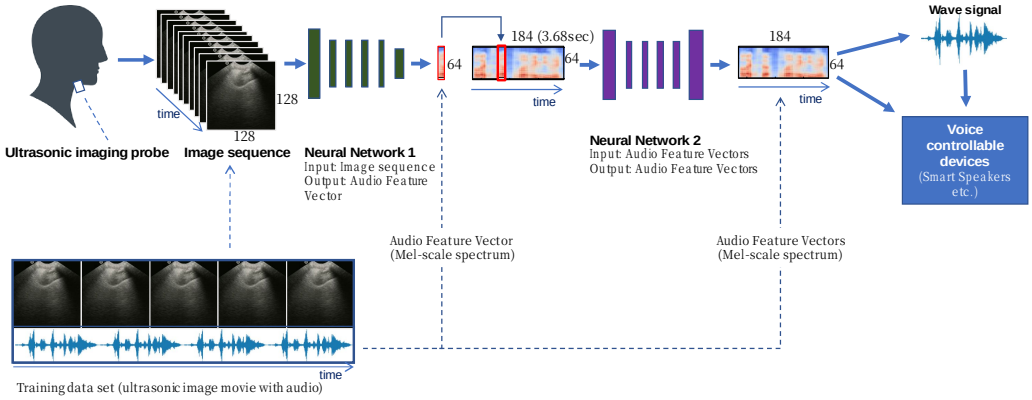
<u>超音波イメージングによる無発声認識</u>

超音波イメージングによる機器構成を右図に示す。顎下のプローブからの映像を実時間でニューラルネットワークに送信し、音声スペクトログラム（MFCC)を出力とする。MFCC を Griffin Lim アルゴリズムによる音声波形に復号し、音声出力を得る。出力された音声波形により、既存のスマートスピーカを駆動できることを確認した。

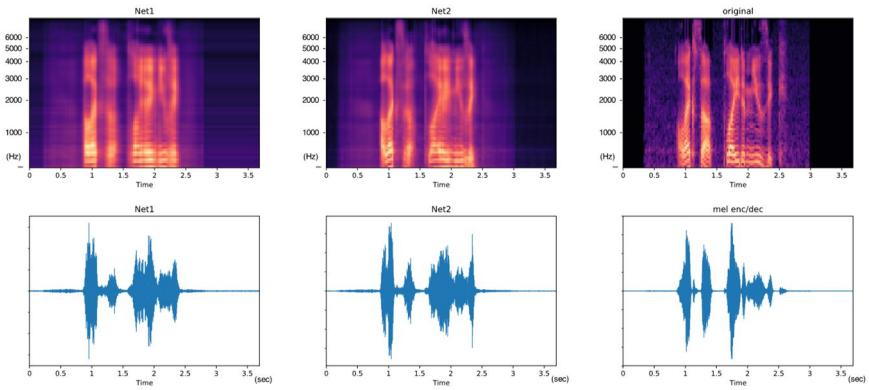




変換に利用したニューラルネットワークの構成を図に示す。Network 1 は，畳み込みニューラルネットワーク(CNN)4 層、{Conv2D - LeakyReLU - Dropout - Batch Normalization}とそれに続く 6 つの層 {Flatten - Dense - LeakyReLU - Dropout - Dense - LeakyReLU}で構成されている．Network1 の出力サイズは、音響特徴ベクトル(すなわち，64)の長さと同じである。入力画像と出力ベクトルのスカラー値は共に 0～1 に正規化されている。損失関数は平均二乗誤差、最適化関数は Adam を使用している。音質を向上させるために，Network 2 は音響特徴ベクトルの列を取り，入力と同じ長さの音響特徴ベクトル列を生成する。



実際に生成された音声波形の例を右図に示す。上段は音響特徴ベクトル(メルスケールスペクトラム)，下段は対応する波形である．Net1 とラベル付けされたグラフは Network1 の結果であり、Net2 は Network1 + Network2 の結果である。「Original」は，学習データとして採取した音声データをメルスケールスペクトラムに変換したものと、それをさらに音声に復元したものであり、学習における ground truth とみなすことができる。認識精度は Network2 で生成された音は Network1 で生成された音よりも優れた品質となることが判明した。

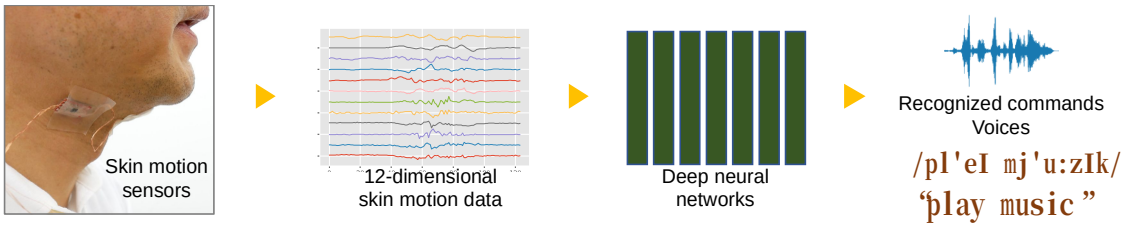


<u>加速度センサによる無発声音声認識</u>

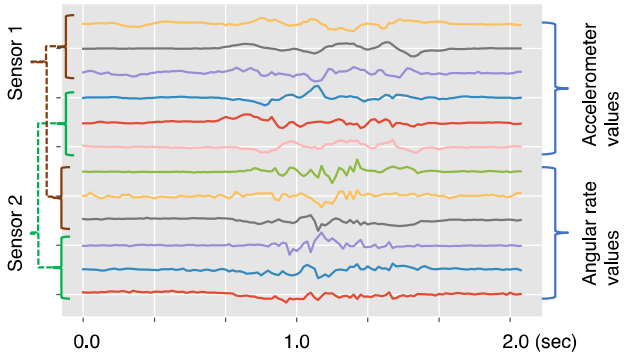
上記の研究により、無発声発話を深層学習で解析することで、音声復元および音声認識を行うことが可能であることが示された。超音波イメージングは、外部から口腔内の情報を得る強力な手段であるが、センサーが高価であること、使用に際してプローブにゲルを塗布する必要があることなどが実用上の課題として残った。そこで、より簡易なセンサー構成で同等の効果を得るための研究を行った。

本研究では，顎下皮膚に装着した MEMS(micro electromechanical systems)加速度計/角速度センサーを使用し，顎運動および舌筋の運動を計測することで無声発話を認識する手法を提案し

ている。顎下に設置された 2 つの MEMS センサーで 12 次元の皮膚運動情報を取得し、深層学習により解析し、35 種類の発声コマンド/フレーズを 94%以上の認識率で識別できることを確認した。また，Connectionist Temporal Classifier（CTC）を用いて、音素記号系列を生成するニューラルネットワークにより、有声発話とは直接対応していない無声発話時の皮膚運動情報から有声発話を生成することを示した。

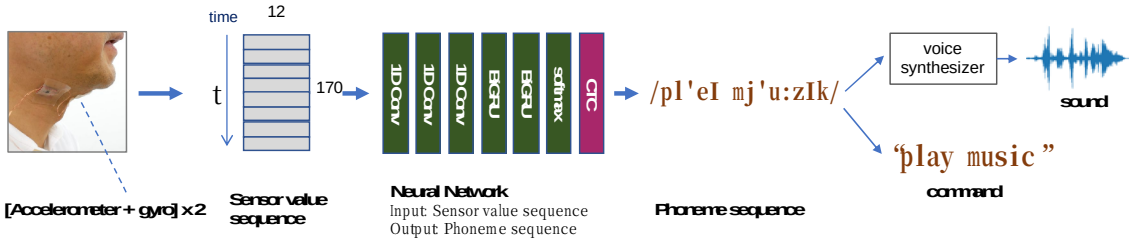


上図にシステムの全体構成を示す。使用するセンサーは，STMicro Electronics 社の LSM9DS1 であり、三軸加速度センサー、三軸角速度（ジャイロ）センサーを搭載している。計測値は 6 自由度（6DOF）となり、2 つのセンサーを使用するので計測値は 12 次元になる。計測されたデータは、I2C を経由してマイクロプロセッサボード Raspberry PI の GPIO ポートに接続される。読み込み処理時間などを含め、検知レートは 58.3 fps であった。図に取得されたセンサー値の例を示す。



無発声発話時のセンサー値から、二種類の方法での復号を行った。一つめは音声コマンドとして復号する方法で、もう一つは発音記号を復号する方法である。それぞれ、ニューラルネットワークとして畳み込み（CNN）層のみを用いる場合と双方向リカレントユニット(GRU)を用いる方法とで精度検証を行った。

本研究のもう一つの貢献は、有声発話と無声発話による学習方法の改善である。前項の超音波イメージングによる手法では、有声発話時に取得した口腔イメージと音声データの組によって、口腔イメージから音声を復元するニューラルネットワークを学習させ、無声発話時に適用していた。しかし、この手法では、有声発話時と無声発話時での口腔運動が同一でないことが認識精度に影響を与えていた。そこで、無声発話のみを学習データとするために、システムが提示した文字列を無声発話で読み上げた場合の口腔運動を記録し、文字列に相当する音素記号を復元する学習手法を構築した（下図）。



これにより、有声発話と無声発話での差分を考慮する必要がなく、口腔運動から音声（音素記号）を復元することが可能になった。音素記号文字列と口腔運動の対応（アラインメント）を考慮するために、音声認識で用いられている CTC 層をニューラルネットに追加している。

右図に、復号された音素記号列の例を示す。正解の音素列には完全に一致しないが、TTS(text to speech)により音声化して復号した場合には聴感上はほぼ同等と聞こた。外部の音声認識機器（スマートスピーカ等）でも認識が可能であることを確認した。

原文	Alexa, volume down.
推定	/a#l’Eks@ v’Oju:m d’aUn/
正解	/a#l’Eks@ v’Olju:m d’aUn/
原文	Alexa, what’s on today?
推定	/a#l’Eks@ w,Ot , On t@d’eI/
正解	/a#l’Eks@ w,Ots ,On t@d’e/
原文	Alexa, set alram for 7am.
推定	/a#l’Eks@ s’Eta#l’A@m f0@ s’Ev@n ,eI’Em/
正解	/a#l’Eks@ s’Et a#l’A@m f0@ s’Ev@n ,eI’EmI/

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　8件／うち国際共著　2件／うちオープンアクセス　0件）

１．著者名 Hirotaka Hiraki, Jun Rekimoto	４．巻 2021
２．論文標題 SilentMask: Mask-type Silent Speech Interface with Measurement of Mouth Movement	５．発行年 2021年
３．雑誌名 Augmented Humans 2020	６．最初と最後の頁 pp.1-8
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 －

１．著者名 Jun Rekimoto, Yu Nishimura	４．巻 2021
２．論文標題 Derma: Silent Speech Interaction Using Transcutaneous Motion Sensing	５．発行年 2021年
３．雑誌名 Augmented Humans 2020	６．最初と最後の頁 pp.1-8
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 －

１．著者名 Kimura Naoki、Hayashi Kentaro、Rekimoto Jun	４．巻 2020
２．論文標題 TieLent: A Casual Neck-Mounted Mouth Capturing Device for Silent Speech Interaction	５．発行年 2020年
３．雑誌名 AVI 2020	６．最初と最後の頁 pp.1-8
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3399715.3399852	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 Zhang Xinlei、Su Zixiong、Rekimoto Jun	４．巻 2020
２．論文標題 Aware: Intuitive Device Activation Using Prosody for Natural Voice Interactions	５．発行年 2022年
３．雑誌名 Proceedings of the ACM on Human-Computer Interaction	６．最初と最後の頁 1-16
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3491102.3517687	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 該当する

1．著者名 Zhang Xinlei、Miyaki Takashi、Rekimoto Jun	4．巻 5
2．論文標題 JustSpeak: Automated, User-Configurable, Interactive Agents for Speech Tutoring	5．発行年 2021年
3．雑誌名 Proceedings of the ACM on Human-Computer Interaction	6．最初と最後の頁 1～24
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3459744	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 該当する

1．著者名 Kawamura Kazuki、Rekimoto Jun	4．巻 2021
2．論文標題 A Language Acquisition Support System that Presents Differences and Distances from Model Speech	5．発行年 2021年
3．雑誌名 Annual ACM Symposium on User Interface Software and Technology	6．最初と最後の頁 44-46
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3474349.3480225	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 Koike Hideki、Rekimoto Jun、Ushiba Junichi、Furuya Shinichi、Ito Asa	4．巻 2021
2．論文標題 Human Augmentation for Skill Acquisition and Skill Transfer	5．発行年 2021年
3．雑誌名 Extended Abstract of the ACM on Human-Computer Interaction	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） 10.1145/3411763.3441354	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 Su Zixiong、Zhang Xinlei、Kimura Naoki、Rekimoto Jun	4．巻 2021
2．論文標題 Gaze+Lip: Rapid, Precise and Expressive Interactions Combining Gaze Input and Silent Speech Commands for Hands-free Smart TV Control	5．発行年 2021年
3．雑誌名 ACM Symposium on Eye Tracking Research and Applications	6．最初と最後の頁 1-6
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3448018.3458011	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　4件／うち国際学会　5件）

１．発表者名
Jun Rekimoto

２．発表標題
Human Augmentation and the future of Human-Computer Integration

３．学会等名
IEEE InTech 2020（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
暦本純一

２．発表標題
Human Augmentation:人間の能力の拡張と進化

３．学会等名
MIRU2020（招待講演）

４．発表年
2020年

１．発表者名
暦本 純一, 木村 直紀, 河野 通就

２．発表標題
SottoVoce: 超音波画像と深層学習による無発声音声インタラクション

３．学会等名
インタラクション2019

４．発表年
2019年

１．発表者名
Jun Rekimoto

２．発表標題
Homo Cyberneticus: The Era of Human-AI Integration

３．学会等名
ACM UIST 2019（国際学会）

４．発表年
2019年

１．発表者名
Jun Rekimoto

２．発表標題
Human Augmentation and the future of Human-Computer Interactions

３．学会等名
CHIuXID, 5th International ACM In-Cooperation HCI and UX Conference（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Jun Rekimoto

２．発表標題
Human Augmentation(keynote)

３．学会等名
ACM MobileHCI2019（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Naoki Kimura, Michinari Kono, Jun Rekimoto

２．発表標題
SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks

３．学会等名
CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems（国際学会）

４．発表年
2019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1件

国際研究集会	開催年 null年

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