Derma: 皮膚運動計測によるサイレントスピーチインタラクション

暦本純一[1,2,a)]　西村 悠[2,b)]

概要：
サイレントスピーチインタラクション (SSI) は，有声ではない発話による音声インタラクション手段であり，ウェアラブルコンピューターなど，さまざまな状況での対話手段として，また発声困難者への支援技術としての可能性を持つ．従来より，画像（リップリーディング），筋電，超音波エコー映像などを利用する手法が提案されていたが，それぞれに制約があった．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する．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たところ，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す．本研究の構成は，従来のサイレントスピーチの認識手段と比較して，装着時に目立たなく，小型軽量であり，照明条件などの周囲の環境要因の影響を受けにくい．

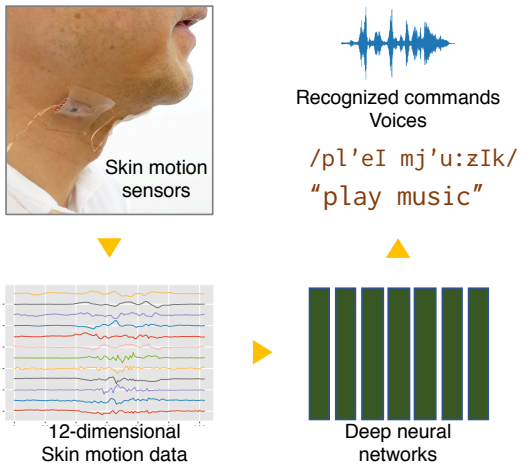

図 1　Derma は，無声発話を顎下の皮膚運動を加速度/角速度センサーから取得されたデータにより認識する．２つの６自由度 (3 軸加速度および 3 軸角速度）センサーを皮膚に貼り付けて使用する．

1. 背景

音声対話システムが多くの状況で使用されるようになってきた．スマートフォン，スマートスピーカー，ウェアラブルコンピューター等のデジタルデバイスを音声で制御できることは多くの可能性を生み出す [32]．音声認識精度の向上により，音声対話はインタラクションの手段として重要かつ現実的なものとなった．音声対話は視覚的な注意を必要とせず，ジェスチャーや GUI などの他の対話手段と併用できる．利用者が他のタスク，たとえば運転，料理，宿題の実行，デスクトップまたはノートブックコンピューターの使用，を行っているときにも利用できる．

しかし，音声インタラクションは二つの課題があると考える．第一に，公共の場での音声対話には制限がある．発声が周囲の人の迷惑となり，個人情報や秘密情報を発声することができない．第二に，ノイズの多い環境では音声認識の精度が低下する．これらの問題は，音声インターフェイスを使用してウェアラブルあるいはモバイルコンピューターと対話しようとする場合に重要な課題となる．

2. サイレントスピーチインタラクション (SSI)

これらの課題に対処するために，声帯を振動させずに無声で発話するサイレントスピーチインタラクション（silent speech interaction, SSI）に関する研究が行われている [12], [14]．SSI は，ユーザーが既に持っている発話能力を利用できる．ジェスチャ入力と比較した場合，新たなジェスチャーコマンドを習得する必要がないなど，音声インタラクションが持つ多くの利点を継承できる．また，声帯損傷などの理由で有声発話が困難な方への支援技術とし

1　東京大学大学院情報学環
2　ソニーコンピュータサイエンス研究所
a)　rekimoto@acm.org
b)　nishimura@csl.sony.co.jp

ての可能性がある．以下に示すような手法が提案・研究されている：

リップリーディング (画像方式): カメラにより話者の口唇あるいは顔全体の画像を取得し，発話内容を推定する [4], [37], [39]． 画像方式では，発話者の顔の前にカメラを設置する必要があり，その形態は，ウェアラブルやモバイルアプリケーションとして適さない． 携帯電話のカメラを使用することも考えられるが [37]，携帯電話を利用者の顔を正しく撮影できる位置に保持する必要があり，機器保持のため手が束縛されてしまう． 画像方式は，光の状態（日光，暗い場所など）の変化によって影響を受ける可能性があり，また一般に電力を消費する．

非可聴つぶやき（Non-audible murmur, NAM）: 利用者の皮膚または喉に装着されたマイクでつぶやき発話を認識する [19], [30]． 利用者は，声帯の振動を伴わない呼吸音 (ささやき声) で発話する． ただし，システムが正確に認識しようとするためには，利用者のささやき声は近隣の人に聞き取られてしまう可能性がある．

超音波イメージング: 超音波イメージング [13] は身体に放射される超音波の反射時間を測定することで身体の内部状態を認識する． これを SSI に適用し，顎の下に取り付けられた超音波プローブにより口腔内の状況を計測し，発話を推定する [11], [20], [22], [38]． ただし，超音波イメージングのためのプローブは非常に複雑で高価であり，体内画像を正確に取得するにはゲルを皮膚に塗布し，プローブを皮膚に接触させる必要がある．

Electromyography (EMG): 筋電図（EMG）を使用して取得される口腔付近の筋肉の状況から音声を推定する [21], [28], [34]． 口腔の動きを使用するジェスチャ認識の一種であるため，検出可能なコマンドの数は限られており，利用者は既存の発話スキルを使用する代わりに新しいジェスチャスキルを習得する必要がある．また顔表面に電極を添付する必要があり，装着が目立つ．

口の中または周辺のデバイス: 口腔内，あるいは周辺にセンサーを配備する方式も提案されている．TongueBoard は，口に置かれた静電センサーアレイで音声と舌のジェスチャーを認識する [25]．センサーは口の中に配置する必要があるため，会話や食事などの日常生活と共存できない．

福本は，マイクを口の前面に非常に近づけたセンサーにより，吸気しながらの発声を取得する方式を提案している [15]．ただし，口の前にデバイスを配置する外見は目立ち，通常の会話などの日常生活との共存に課題がある．

唇と舌に取り付けられた小型磁石による手法も提案されている．磁石の動きによる磁場の変化を感知し，発声を推定する [16]．磁石の埋め込みに手術が必要であり，発話困難者の補助技術に用途が限定されている．

これらの方法は，SSI の興味深い発展方向を示しているが課題も多い．インプラントは，持続的なインタラクションを達成する可能性があるが手術を要し利用範囲が限定される．また，提案されているセンサー構成は目立ち，日常生活との共存が困難である．

図 2 “Tadoma”(触診リップリーディング) 手法 (Photo: Courtesy of Perkins School for the Blind Archives) [33]

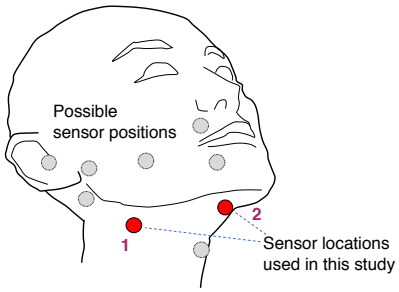

図 3 センサー設置候補箇所：赤で示す二箇所が本研究で使用する箇所

上記の課題を解決するために，本研究では，顎と喉の周りの皮膚運動計測に基づく Derma と呼ばれるサイレントスピーチ認識方法を提案する．

3. Derma:皮膚運動計測による SSI

Derma は，皮膚運動計測に基づくサイレント音声認識技術である．私たちが（無声あるいは有声で）発話するとき，下顎から喉までの皮膚が変動する．これは，舌を制御する舌筋が顎裏にあること，また口の開閉そのもので下顎が変動することによる．この変動を感知して，発話の内容を推定することが本研究の基本的な発想である．顎下の皮膚に MEMS (micro electromechanical systems) 加速度/角速度センサーを装着し，皮膚の動きを測定する．計測情報を深層学習で分析し，発話内容の推定を試みる．

MEMS センサーは，カメラや超音波エコー用プローブと比較すると，非常に低い電力消費で動作し，小型軽量である．画像方式とは異なり，センサーは顎下などの目立たない位置に装着でき，利用者の顔の前に配置する必要がなく，光や環境ノイズなどの外部環境の変動の影響を受けない．

口の周りの皮膚の変形と動きから発話内容を読み取るという発想は，視聴覚障害者のための Tadoma 法（「触診リップリーディング」） [3], [31] に遡ることができる．Tadoma 法では，話者の唇や顎周辺を指で触れ，発話を読み取ろうとする（図 2）．本研究の方法は，Tadoma 法をセンサーと機械学習により自動化する試みであるとも言える．

Tadoma 法では口唇付近にも指を接触させているが，本

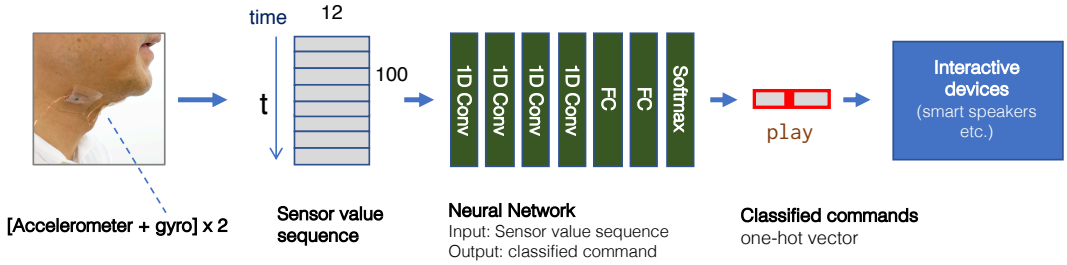

図 4　Derma システム構成

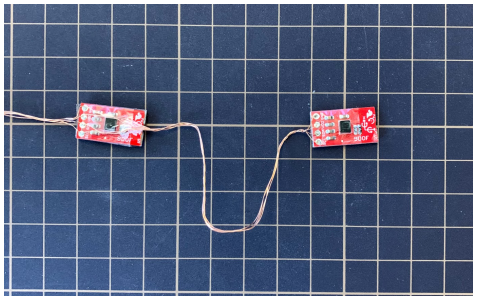

図 5　実験に用いる加速度・角速度センサー (STMicroelectronics 社 LSM9DS1).

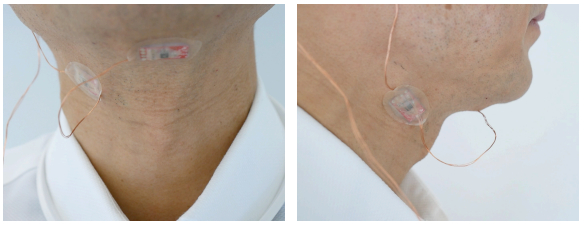

図 6　センサー装着箇所

研究では，常時装着を想定し，センサーを目立ず日常生活を阻害しない位置に配置することを試みる（図 3）．1 つは顎に，もう 1 つは顎下の喉付近に配置した．これらの場所は，口の開閉に関わる下顎の運動，および，舌の運動に関わる舌筋の運動を捉えることができると想定している．将来的には，センサーを他の場所に配置した場合，センサーの数を増やすあるいは減らした場合の性能も調査する予定である．

3.1　センサー構成

図 4 にシステムの全体構成を示す．使用するセンサーは，STMicroelectronics 社の LSM9DS1 [36] 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，および三軸磁力計が $3.5 \times 3 \times 1.0mm$ のパッケージに含まれている（図 5 ）．加速度と角速度情報を使用し，磁力計の値は今回は用いない．したがって計測値は 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（$6DOF \times 2$）になる．計測されたデータは，I2C を経由して 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 6 にセンサー装着位置を示す．

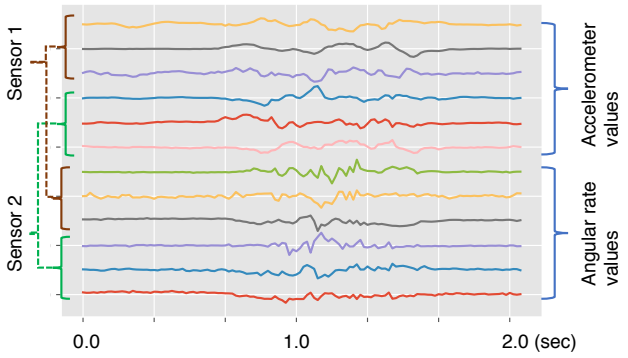

図 7　取得されたセンサー値系列の例：無声で “play music” と発話した場合の正規化されたセンサー値を示す．

表 1　識別実験に持ちた 21 種類のコマンド (発話しない場合も識別対象となるので，識別クラス数は 22 となる).

music	cancel	answer
yes	menu	Alexa
no	open	mute
start	close	left
stop	home	right
play	next	play music
ok	back	stop music

加速度センサーは重力の影響を受け，運動がない場合でも計測値はゼロにならない．これを補正するために，過去の一定期間のセンサー取得値により平均値を求め，原点がゼロになるように調整している．加速度値，角速度値ともに，それぞれの標準偏差により値を正規化している．図 7 に，利用者が “play music” と無声発話したときのセンサー値を示す．

4.　音声コマンドの識別

このセンサー用いて，ボイスコマンドの識別を試みた．さまざまなデジタルデバイスを操作するために使用される 21 種類の音声コマンドを分類対象として使用する（表 1）．

音声コマンドを識別するために，2 種類のニューラルネットワーク（`net-GRU` および `net-Conv`）を評価した．ネットワークへの入力は，固定長（$n = 100$, 1.73 秒に対応）の 12 次元の正規化されたセンサーデータで，出力はコマンド分類用の 22 次元 (21 個のコマンド及び無発話) の one-hot-vector である．

表 2 net-GRU のネットワーク構成. N は識別クラス数 (= 22)

Layer	Size / Stride / Pad	Input dimension
Conv1D	3/1/valid	100x12
Dropout		98x64
Conv1D	3/1/valid	98x64
Dropout		96x128
Conv1D	3/1/valid	96x128
Dropout		94x256
Conv1D	3/1/valid	94x256
Dropout		92x512
BiGRU		92x512
BIGRU		92x512
Dropout		512
FC	4096	512
FC-softmax	N	4096

表 3 net-Conv のネットワーク構成．N は識別クラス数 (= 22)

Layer	Size / Stride / Pad	Input dimension
Conv1D	3/1/valid	100x12
Pooling	2/none/valid	98x64
Dropout		49x64
Conv1D	3/1/valid	49x64
Pooling	2/None/None	47x128
Dropout		23x128
Conv1D	3/1/valid	23x128
Pooling	2/None/valid	21x256
Dropout		10x256
Conv1D	3/1/valid	10x256
Pooling	2/None/valid	8x512
Dropout		4x512
FC	4096	2048
FC-softmax	N	4096

net-GRU は，画像による SSI システムである Lip-Interact [37] で提案されたニューラルネットワークを参考に 2D Conv を 1D Conv に置き換えたものである．入力は 1Dconv, Maxpooling, Dropout からなる層を 4 回繰り返し，双方向リカレントユニット (BiGRU [7]) の 2 つの層によって処理される．BiGRU の出力は 2 つの全結合層によって処理され，最終層の出力が softmax アクティベーション関数により分類コマンドとなる．ネットワーク構成とハイパーパラメタを表 2 に示す．

もう一つのニューラルネットワーク構成 net-Conv は，GRU を用いず，畳み込みのみに基づいている． 1DConv，Maxpooling，Dropout からなる層を 4 回繰り返し，その後に全結合層を接続する．このネットワークでは，各畳み込み層の後に時間方向の長さが短くなっていく． 100×12 の入力は，畳み込み層の後で 8×512 となる．次に，このデータは Dense 層を経て softmax アクティベーション関数により音声コマンドを特定する one-hot-vector となる．ネットワーク構成とハイパーパラメタを表 3 に示す．

これらのネットワークモデルを，Tensorflow [2] をバックエンドとし，NVIDIA TITAN RTX GPU ボードを備えた Keras [8] 深層学習プラットフォームに基づいて実装した．

表 4 無声発話からのコマンド認識結果 (Commands: コマンド数, Average: 個別各実験参加者ごとの結果の平均値, Mixed: 全参加者のデータから学習した場合の結果, ⋆: センサー一個で学習した場合の結果，◇：センサーを貼り直した場合の結果)

Net type	Commands	Average	Mixed
net-GRU	21	91.90%	86.94%
net-Conv	21	**98.93%**	97.01%
net-GRU	35	91.15%	89.27%
net-Conv	35	**98.28%**	97.40%
net-Conv⋆	21	**97.50%**	94.02%
net-Conv◇	21	**91.68%**	
net-Conv◇	35	**94.49%**	

4.1 評価

構築した認識手法を評価するために，5 人の実験参加者（女性 1 人，男性 4 人）による評価実験を行った．各参加者は，無発声を含む 22 種類の音声コマンドを 40 回発声し，データを記録する． 1 セッションは 10 回の有声発話と 10 回の無声発話からなり，参加者は 2 つのセッションを実施する．セッション中にセンサーデータと音声波形を記録する．2 つのセッションの間で，センサーを参加者の喉から取り外し，再度装着し，センサーの取り付け位置のわずかな変位の影響を取り入れるようにする．入力を取り出すための時間位置は，標準偏差 10 の正規分布乱数によって変更する．これは，データ拡張として機能することが期待される．

最終的に，各被験者から 880 回（$= 22 \times 40$）の発話データ，5 人の参加者全員から合計 4,400 回の発話データを収集した．このデータからランダムに選択した 80%を訓練に，残りの 20%を検証に使用する．

評価結果を表 4 に示す． 「平均」列は各実験参加者のデータごとに個別にトレーニングした場合の分類精度の平均を示し，「混合」列はすべての参加者のデータを使用した分類精度を示す． 表に示すように，net-Conv は，net-GRU よりも分類精度が優れている． 各参加者ごとに訓練した場合，98%以上の精度をコマンド分類が達成できる． 図 8 および図 9 に混合行列を示す．

また，net-Conv の学習速度は net-GRU の学習速度よりも大幅に高速であることを確認した． net-GRU は 1 エポックの学習に 13 秒要するが，net-Conv は 1 秒であった． すなわち，net-Conv は net-GRU よりも 13 倍速く学習する．

4.2 コマンド数を増やした場合の評価

さらに，コマンド数を増やした場合の分類性能を評価するために，元の 21 個のコマンドに 15 個のフレーズ（“Alexa, play jazz”, “Alexa，what’s the weather like?” など）を追加した． 表 4 に結果を示す． 追加したような長いフレーズはより識別しやすいということに留意すべきではある

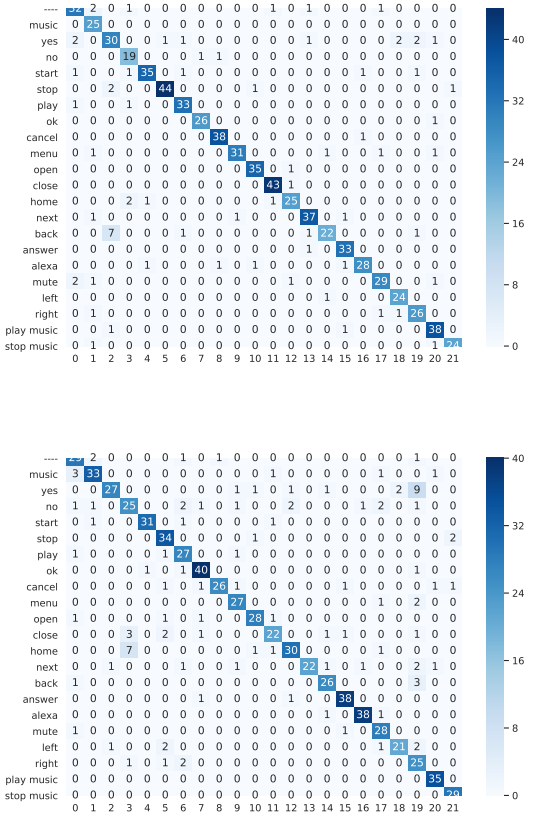

図 8　net-GRU の混合行列: 上: 各実験参加者ごとに訓練した場合; 下: 全参加者のデータから訓練した場合

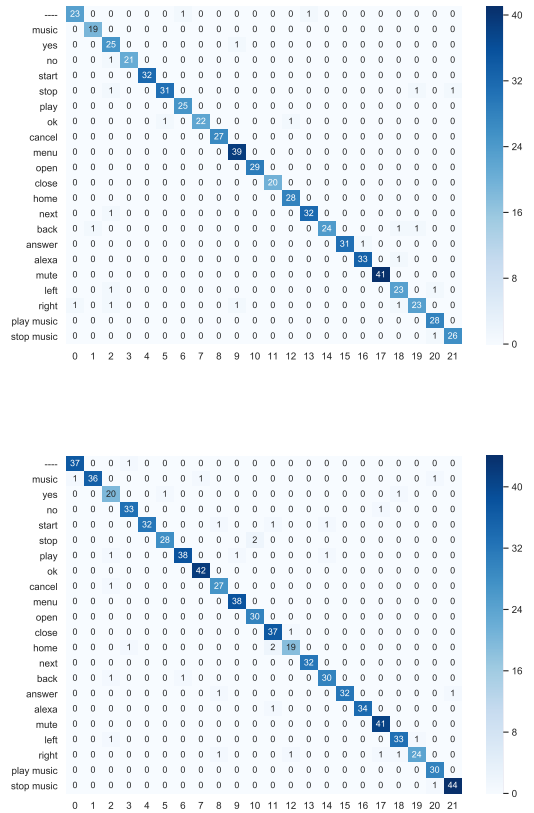

図 9　net-Conv の混合行列: 上: 各実験参加者ごとに訓練した場合; 下: 全参加者のデータから訓練した場合

が，分類精度は以前の結果とほぼ同等であった (表 4 のコマンド数 35 で示した欄).

4.3　センサーを貼り直した場合の評価

上記の実験では，計測途中にセンサーを貼り直しているが，学習データと検証データ間で，同じセンサーの貼付状況での計測値が一部使われていた．実際の利用を想定すると，センサーを完全に貼り直しても認識性能が維持できることが望ましい．そこで，同一の被験者を対象に，上記の実験とは異なる日にセンサーを改めて貼付し，検証用データを取得した．これを用いた認識精度は，21 コマンドの場合で 91.68%，35 コマンドの場合で 94.49% であった (表 4 の ◇ で示した欄).

以上の評価実験から，net-Conv が精度と学習速度の点で優れており，センサーの再貼付をしない場合には 98%以上，再貼付をした場合には 94%以上の精度で発話コマンドが認識できることが確認できた．

これにより，提案方式による無声対話の可能性が示された．今回の実験では，実験参加者は着座した状態で学習データを収集したが，歩きながらなど，他の運動が加味されている状況での精度検証を今後進める予定である．また，センサーを再度貼付し直した場合にセンサーの位置が変化しないような貼付手法についても検討する予定である．

5.　Connectionist Temporal Classification (CTC) による発話音声生成

前章では，登録されたコマンドを識別するためのネットワークを評価した．この場合，無声発話から音声を聞き取ることはしないので，コマンドが誤認識された場合，利用者にとっては自己の発話を改善する手がかりがない．一方，SSI システムが，認識した無声発話から音声を生成する場合，利用者はそれを聞き取ってシステムにとってより認識しやすく改善する方法についてフィードバックを受け取ることができる．また，無声発話からの音声合成が可能な場合，近くにある音声認識機能を備えたデジタル機器（スマートスピーカーなど）を無改造で操作できる．さらに，声帯損傷などの理由により有声発話が困難な利用者のためのコミュニケーション支援技術の利用価値がある．この場合，合成された音声は，デバイスではなく人間界の対話支援に用いられる．

無声発声を音響に変換するための学習方法として，次の二種類の方式を考える．

第一の **half-silent 法**では，有声発声時に音声データとセンサーデータを収集する．これを教師データとし，センサーデータから音声を再現するように学習を行う．多くのリップリーディング研究では，この方法を採用し，音声つ

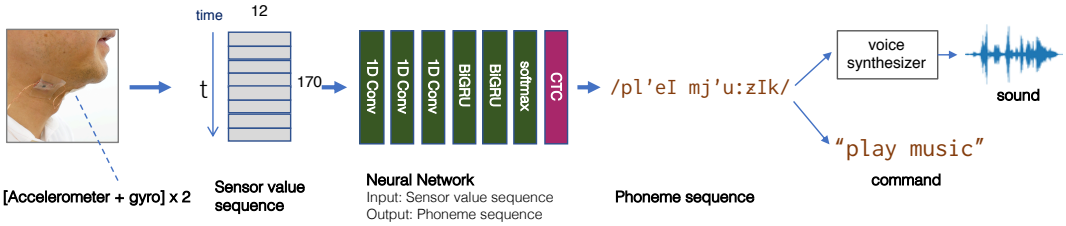

図 10　音素記号を生成する場合のシステム構成

きのビデオを学習し，映像のみから音声を再現することを試みている．ただし，この方法では，無声発話時と有声発話時とで口腔が同じように動くことを前提としている．さらに，有声発話が困難な人や聴覚に障害がある人からは，訓練データを得ることができない．そこで，2 番目の方式である **pure-silent 法**では，利用者は無声発話のみを実行し訓練データを得る．

5.1　Pure-silent 法

pure-silent 法の場合，有声発話は利用できないので，学習データの収集方法を検討する必要がある．本研究では，次の手順でトレーニングデータを収集する方式を提案する：

- システムは読み上げるべきテキストを利用者に提示し，利用者は無声発話でそれを読み上げる．
- 提示されたテキストを音素列に変換し，無声発話時のセンサーデータとともに記録する．
- 収集したデータから，システムは無声発話時のセンサーデータから音素系列への変換を学習する．
- 翻訳された音素系列を，音声合成装置で実際の音声に変換する．

この方法では，利用者の元の音声は再現されないが，音声コミュニケーションには使用できる． GRID [9] などの既存の読唇コーパスは，有声発話のデータのみを収集するが，この方法は，無声発話のみから学習できる点に独自性がある．

有声発話中の口の動きを記録する従来の方法と比較して，無声発話特有の発話方法にも適用できる．たとえば，口唇を閉じた状態での発話からの音声再現にも適用できる．

ただし，ここでの課題は，無声発話中の，各音素記号に対応するタイミングを，テキストから変換された音素列との時間対応（アラインメント）を達成することである．これを手作業で行うことは極めて困難で労力を要するが，多くの音声認識およびテキスト読み上げニューラルネットワークシステムおよびその他の無音音声システム [5] で効果的であると報告されているコネクショニスト時間分類（CTC） [17] を使用することで達成する．

コネクショニスト時間分類（Connectionist Temporal Classification, CTC）は，ニューラルネットワークの訓練のために用いられる手法で，リカレントニューラルネットワーク（RNN）（LSTM，GRU など）を学習させる場合にアラインメント情報を明示的に必要としない手法である．

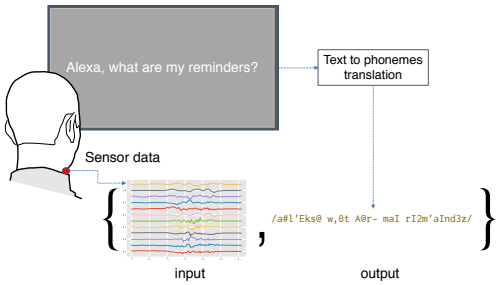

図 11　"pure-silent" 法の学習データ収集方法

CTC はオンライン手書き認識，音声合成，音声認識などに使用されている．このようなシステムを学習させる際に，入力と出力の各要素の対応が不明な点が課題となる．例えば音声認識では，観測された計測値 (音響特徴) の系列と，ターゲットラベル（音素列）間の対応をもった訓練データを準備することは通常困難である．そのため，各タイムステップでアラインメントの確率分布を予測する必要がある．

CTC では，ブランクあるいはターゲットのシンボルを複数回繰り返すことにより，アラインメントのさまざまな可能性を表現する．たとえば，"Hello" に対応する発話の場合，ターゲットの音素系列は`/h'EloU/`になる．この系列を`/h'E_looU/`，`/_h'Elo_UU/`，または`/h'_EEEloUU/`などのさまざまな時間アライメントに展開する (`/_/`はブランク)．CTC は，これらの生成された音素列とニューラルネットワークの出力と一致させる可能性を計算する損失関数を提供する．

5.2　ネットワーク構成

上記の手法に基づいたシステム構成を図 10 に示す． 前節で説明したシステムと同様に，利用者の無声発話から得られた加速度/角速度センサーデータが入力となるが，出力は音素系列である． 音声合成装置を用いて音素系列を実際の音声に変換する．

表 5 にニューラルネットワークの構成とハイパーパラメタを示す．CTC を除き，前節の `net-GRU` とほぼ同じである． 3 つの BiGRU 層が接続され，3 つの Conv1D 層が続く． 出力が CTC で未定長の音素シーケンスと比較される． CTC により，入力センサー系列と出力音素シーケンスの時間方向のアラインメントを明示的にとる必要がない． 実装には，CTC の Keras 実装である CTCModel [35] と， テキストを音素に変換し，音素から音声を合成するた

Layer	Size / Stride / Pad	Input dimension
Conv1D	3/1/same	170x12
Dropout		170x64
Conv1D	3/1/same	170x64
Dropout		170x256
Conv1D	3/1/same	170x256
Dropout		170x512
BiGRU		170x512
BIGRU		170x512
BIGRU		170x512
FC-softmax	170	170

表 5 音素列生成のためのネットワーク構成

Alexa, play music.
`/a#l'Eks@ pl'eI mj'u:zIk/`
Alexa, play jazz.
`/a#l'Eks@ pl'eI dZ'az/`
Alexa, stop music.
`/a#l'Eks@ st'0p mj'u:zIk/`
Alexa, what's on today.
`/a#l'Eks@ w,0ts ,0n t@d'eI/`
Alexa, set alarm for 7 am.
`/a#l'Eks@ s'Et a#l'A@m f0@ s'Ev@n ,eI'Em/`
Alexa, set timer for 10 minutes.
`/a#l'Eks@ s'Et t'aIm3 f0@ t'En m'InIts/`
Alexa, what are my reminders?
`/a#l'Eks@ w,0t A@r- maI rI2m'aInd3z/`

図 12 テキストと対応する音素列の例

表 6 音素生成の単語エラー率 (WER) および文字エラー率 (CER) による精度評価 (P1–P4: 実験参加者ごとの結果, Average: 全参加者の結果の平均)

	P1	P2	P3	P4	Average
WER	0.075	0.108	0.156	0.024	**0.091**
CER	0.041	0.056	0.088	0.07	**0.048**

めに，ESpeakNG [1] 用いる．音素の表記は，ESpeakNG で定義されているものを用いている．

5.3 評価

4 人の実験参加者による評価実験を行った． 各参加者は，画面に表示されるフレーズを無声発話により発話する．取得したセンサーデータと，音素記号をトレーニングデータとして保存する（図 11）． 図 12 は，サンプルテキストフレーズとその音声記号表現を示している． 35 個のフレーズを，各参加者は無声で 20 回，有声で 20 回読み上げる．

表 6 に，生成した音素列とターゲットの音素列を比較して得られた単語エラー率（word error rate, WER）および文字エラー率（character error rate, CER）による結果を示す． また，ESpeakNG 音声シンセサイザーを使用して，出力音声フレーズを実際の音声に変換し，生成された音声なしで既存のスマートスピーカー（Amazon Echo）を制御

原文	Alexa, volume down.
推定	`/a#l'Eks@ v'0ju:m d'aUn/`
正解	`/a#l'Eks@ v'0lju:m d'aUn/`
原文	Alexa, what's on today?
推定	`/a#l'Eks@ w,0t , 0n t@d'eI/`
正解	`/a#l'Eks@ w,0ts ,0n 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/`
原文	Alexa, what's the weather like?
推定	`/a#l'Eks@ w,0ts D@ w'ED3 lIk/`
正解	`/a#l'Eks@ w,0ts D@ w'ED3 l'aIk/`

図 13 近似生成の例 (正解の音素列とは完全には一致しないが，聴感上，あるいは音声認識システムでは正しく認識される)

表 7 音響特徴量を直接生成するネットワークの構成

Layer	Size / Stride / Pad	Input dimension
Conv1D	3/1/valid	24x12
Pooling	2/none/valid	22x12
Dropout		11x64
Conv1D	3/1/valid	11x64
Pooling	2/None/None	9x128
Dropout		4x128
Conv1D	3/1/valid	4x128
Pooling	2/None/valid	2x256
Dropout		1x256
FC	4096	256
FC	20	4096

できることを確認した．

WER は，単語の一致度から計算されるので，音素が 1 文字でも一致しない場合，それは不一致として計算される．実際には，類似の音素を出力する傾向があり，聴感上はほぼ正しく知覚される場合が多かった． このような近似出力の例を図 13 に示す．

以上から，我々の方法で認識されたサイレント音声が，周囲の音声認識デバイスの操作などに使用できる可能性が示された．

6. 音声特徴量の直接生成

最後に，*half-silent* 方式によりセンサー情報を音響情報に直接変換することを試みた．この実験では，利用者が有声音声によって例文を読み上げ，音声データとセンサーデータが記録し，学習データとして使用する．

この場合，ニューラルネットワークを，コマンドを認識する代わりに音響情報を推定するように学習させる．このために，固定長の一連のセンサーデータ $X_t = x_{t-n}, ..., x_{t-1}, x_t$ がネットワークの入力として使用する（現在は $n = 24$ で，400 ms のセンサー情報に相当する），出力は 1 つの音声特徴ベクトル Y_t である．これは 20 次元データであり，基本周波数 F_0 と音声から取得したメル周波数ケプストラム係

数（MFCC）の最初の 19 次元の組み合わせたものである．この音響特徴量を，Griffin-Lim アルゴリズム [18] を使用して音声波形に変換する．

表 7 にネットワーク構成とハイパーパラメタを示す．この構成は，超音波画像を使用するサイレント音声対話システムである SottoVoce で使用されるニューラルネットワーク構成に基づいている [22]．

有声音とセンサーデータを組み合わせて学習したニューラルネットワークにセンサーデータのみを与えると，生成された音が正しく再現された．しかし，無声発話と同時に取得されたセンサーデータから音声を生成しようと試みたところ，音声は正しく再生されなかった．このことは，無声発話と有声発話において，皮膚運動が微妙に異なっている（口腔内の運動も異なっている）ことを示唆しているが，センサーデータの何が異なるのかをさらに調査する必要があると考えている．

7. 議論

サイレントスピーチ認識に要する情報量

本研究の貢献のひとつは，従来の画像ベース等の SSI よりも大幅に単純化されたセンサーで，12 の低い次元でもサイレント音声コマンドを識別できることを証明したことにある． たとえば，30 fps での 128×128 グレースケール画像の情報量は 3,932,160 bps だが，本研究でのセンサー情報量は 5,568 bps である． このような少量の情報でも音声を認識できる点が興味深いと考えている．

人間の音声には多くの冗長性があり，実際の情報量は 39.15 bps [10] である． したがって，これまで考えられていたよりもはるかに単純なセンサー構成で，無声発話が認識できる可能性がある．

センサー数・配置

本研究では，皮膚に取り付けるセンサー数を，喉の下の目立たない 2 箇所に限定した．他の取り付け位置，センサーの数，および認識精度の関係は，今後さらに評価する必要がある．

センサー数を増やし，唇の近くにもセンサーを設置した場合，認識精度をさらに改善できるかが一つの方向である．この場合，センサーの設置がより目立つようにはなるが，発声障害のある人のための支援技術としての利用可能性がある

一方，センサー数を限定し，センサーを目立たないように設置したり，イヤフォンと統合したりした場合に，認識精度と認識可能な語彙数をどれだけ確保できるかも検討課題である．たとえば，イヤホンをタップするなどのコマンドで操作することを想定すると，利用者は新たにタップコマンドを憶えなければならない．サイレント発声コマンドであれば，play, stop など，利用者が知っている単語をそのまま利用できるので，コマンド数が増やすことも容易である．その可能性を検証するために，上記の実験で使用したのと同じデータセットを使用して，単一のセンサー（図 3 で「1」とマークされた場所）のみを使用して `net-Conv` を訓練した結果を表 4 の $\star$ の欄に示す．認識精度は 97.50% で，2 つのセンサーを使用した場合よりわずかに低下したが，それでもかなり良好である．

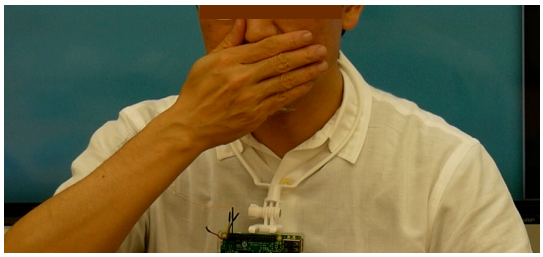

図 14 秘匿発話のために口元を隠しても認識が可能

秘匿性

サイレントスピーチは一般に発話の秘匿性を担保すると考えられているが，画像ベース読唇技術が進歩すれば，発話中の口の動きが撮影できれば発話内容が推定されてしまう．画像ベースのサイレントスピーチ認識手法を使用する場合，口元を隠すことができないのでセキュリティリスクになる． 本研究の手法は口唇映像は必要ないので，利用者は口を手で覆うだけで発話を読唇から隠すことができる（図 14）．

ウェアラブルエレクトロニクス/非接触センサーとの融合

ウェアラブルスキンエレクトロニクスに関する多くの研究があり，長期間皮膚に接着することができ継続的に身体情報を計測することが可能になっている [29]．ウェアラブルエレクトロニクスを本研究の方式に適用することで，センサーの再貼付による認識率の低減を回避できる． Lee らは，音声認識のために喉に取り付けられる極薄の振動応答性素子を開発した [23]． そのような技術と本研究の方法と組み合わせることができれば，センサー部分非常に目立たなくなり，最終的には肌の一部として機能するようになる．

また，非接触で皮膚運動を計測する可能性もある．Time-of-Flight (TOF) による深度センシングや，Soli [26] などのミリ波レーダーによる測定を用いて，喉下に配置されてはいるが皮膚とは非接触なセンサーにより，本研究と同様の効果を達成できるかは興味深い将来課題である．

他の計測手段の併用

他のセンシング方法を組み合わせることも，多くの可能性を有している． 筋電 (Electromyogram, EMG) [21], [28], [34]，耳介内光電式容積脈波 (in-ear photoplethysmography) [24]，耳介内顎運動センシング [6]，および耳介内静電計測 [27] は，このような組み合わせの候補である．

8. 結論

顎下皮膚に付着して皮膚運動を測定する 2 つの加速度/

角速度センサーにより，無声発話を認識する手法を提案した． 2 つのセンサーから得られる 12 次元の皮膚運動情報を深層学習により解析し，35 種類の音声コマンドを識別する実験の結果 94%以上の正答率を得た． また，音素記号を生成する CTC を用いたネットワークにより，無声発声から有声発声への変換を試みた． 従来の SSI 手法と比較して，提案方式は装着が目立たず小型軽量であり，食事や通常の会話などを阻害しにくい．また照明条件などの周囲の環境要因の影響を受けにくい．この特徴から，モバイルでウェアラブルなインタラクションでの利用可能性がある．

参考文献

[1] 2016.
eSpeak NG Text-to-Speech.
(2016).
github.com/espeak-ng/espeak-ng

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015.
TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.
(2015).
https://www.tensorflow.org/
Software available from tensorflow.org.

[3] Sophie Alcorn. 1932.
The Tadoma Method.
Volta Rev. 34 (1932), 195–198.

[4] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. 2016a.
LipNet: End-to-End Sentence-level Lipreading.
(2016).

[5] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. 2016b.
LipNet: End-to-End Sentence-level Lipreading.
(2016).

[6] Abdelkareem Bedri, David Byrd, Peter Presti, Himanshu Sahni, Zehua Gue, and Thad Starner. 2015.
Stick It in Your Ear: Building an In-ear Jaw Movement Sensor. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 1333–1338.
DOI:http://dx.doi.org/10.1145/2800835.2807933

[7] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014.
Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.
ArXiv abs/1406.1078 (2014).

[8] François Chollet and others. 2015.
Keras.
https://keras.io. (2015).

[9] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006.
An audio-visual corpus for speech perception and automatic speech recognition.
The Journal of the Acoustical Society of America 120, 5 (2006), 2421–2424.
DOI:http://dx.doi.org/10.1121/1.2229005

[10] Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019.
Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche.
Science Advances 5, 9 (2019).
DOI:http://dx.doi.org/10.1126/sciadv.aaw2594

[11] Tamás Gábor Csapó, Tamás Grósz, Gábor Gosztolya, László Tóth, and Alexandra Markó. 2017.
DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface. In *INTERSPEECH*.

[12] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg. 2010.
Silent Speech Interfaces.
Speech Commun. 52, 4 (April 2010), 270–287.
DOI:http://dx.doi.org/10.1016/j.specom.2009.08.002

[13] Roman Gr. Maev (ed.). 2013.
Advances in Acoustic Microscopy and High Resolution Imaging: From Principles to Applications.
Wiley-VCH.

[14] Joo Freitas, Antnio Teixeira, Miguel Sales Dias, and Samuel Silva. 2016.
An Introduction to Silent Speech Interfaces (1st ed.).
Springer Publishing Company, Incorporated.

[15] Masaaki Fukumoto. 2018.
SilentVoice: Unnoticeable Voice Input by Ingressive Speech. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. ACM, New York, NY, USA, 237–246.
DOI:http://dx.doi.org/10.1145/3242587.3242603

[16] Jose A. Gonzalez, Lam A. Cheah, Angel M. Gomez, Phil D. Green, James M. Gilbert, Stephen R. Ell, Roger K. Moore, and Ed Holdsworth. 2017.
Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning.
IEEE/ACM Trans. Audio, Speech and Lang. Proc. 25, 12 (Dec. 2017), 2362–2374.
DOI:http://dx.doi.org/10.1109/TASLP.2017.2757263

[17] Alex Graves, Santiago Fernández, and Faustino Gomez. 2006.
Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *In Proceedings of the International Conference on Machine Learning, ICML 2006*. 369–376.

[18] D. Griffin and Jae Lim. 1984.
Signal estimation from modified short-time Fourier transform.
IEEE Transactions on Acoustics, Speech, and Signal Processing 32, 2 (April 1984), 236–243.
DOI:http://dx.doi.org/10.1109/TASSP.1984.1164317

[19] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. 2010.
Silent-speech enhancement using body-conducted vocal-tract resonance signals.
Speech Communication 52, 4 (2010), 301 – 313.

[20] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone.

2010.
Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips.
Speech Commun. 52, 4 (April 2010), 288–300.
DOI:http://dx.doi.org/10.1016/j.specom.2009.11.004

[21] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018.
AlterEgo: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 43–53.
DOI:http://dx.doi.org/10.1145/3172944.3172977

[22] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019.
SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 146, 11 pages.
DOI:http://dx.doi.org/10.1145/3290605.3300376

[23] Siyoung Lee, Junsoo Kim, Inyeol Yun, Geun Yeol Bae, Daegun Kim, Sangsik Park, Il-Min Yi, Wonkyu Moon, Yoonyoung Chung, and Kilwon Cho. 2019.
An ultrathin conformable vibration-responsive electronic skin for quantitative vocal recognition.
Nature Communications 10, 1 (2019), 2468.
DOI:http://dx.doi.org/10.1038/s41467-019-10465-w

[24] Cheng-Lun Li, Ayse G. Buyuktur, David K. Hutchful, Natasha B. Sant, and Satyendra K. Nainwal. 2008.
Portalis: using competitive online interactions to support aid initiatives for the homeless. In *CHI '08 extended abstracts on Human factors in computing systems*. ACM, New York, NY, USA, 3873–3878.
DOI:http://dx.doi.org/10.1145/1358628.1358946

[25] Richard Li, Jason Wu, and Thad Starner. 2019.
TongueBoard: An Oral Interface for Subtle Input. In *Proceedings of the 10th Augmented Human International Conference 2019 (AH2019)*. ACM, New York, NY, USA, Article 1, 9 pages.
DOI:http://dx.doi.org/10.1145/3311823.3311831

[26] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016.
Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar.
ACM Trans. Graph. 35, 4, Article 142 (July 2016), 19 pages.
DOI:http://dx.doi.org/10.1145/2897824.2925953

[27] Denys J. C. Matthies, Bernhard A. Strecker, and Bodo Urban. 2017.
EarFieldSensing: A Novel In-Ear Electric Field Sensing to Enrich Wearable Gesture Input Through Facial Expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1911–1922.
DOI:http://dx.doi.org/10.1145/3025453.3025692

[28] Geoffrey S. Meltzner, James T. Heaton, Yunbin Deng, Gianluca De Luca, Serge H. Roy, and Joshua C. Kline. 2018.
Development of sEMG sensors and algorithms for silent speech recognition.
Journal of neural engineering 15 4 (2018), 046031.

[29] Akihito Miyamoto, Sungwon Lee, Nawalage Florence Cooray, Sunghoon Lee, Mami Mori, Naoji Matsuhisa, Hanbit Jin, Leona Yoda, Tomoyuki Yokota, Akira Itoh, Masaki Sekino, Hiroshi Kawasaki, Tamotsu Ebihara, Masayuki Amagai, and Takao Someya. 2017.
Inflammation-free, gas-permeable, lightweight, stretchable on-skin electronics with nanomeshes.
Nature Nanotechnology 12 (17 Jul 2017), 907 EP –.
https://doi.org/10.1038/nnano.2017.125
Article.

[30] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2003.
Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*., Vol. 5. V–708.
DOI:http://dx.doi.org/10.1109/ICASSP.2003.1200069

[31] C. M. Reed. 1996.
The implications of the Tadoma method of speechreading for spoken language processing. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Vol. 3. 1489–1492 vol.3.
DOI:http://dx.doi.org/10.1109/ICSLP.1996.607898

[32] Alexander I. Rudnicky. 1989.
The Design of Voice-driven Interfaces. In *Proceedings of the Workshop on Speech and Natural Language (HLT '89)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 120–124.
DOI:http://dx.doi.org/10.3115/100964.100972

[33] Perkins School. 1961.
Perkins School for the Blind.
The Lantern 33, 5 (1961), 21.

[34] Tanja Schultz. 2010.
ICCHP Keynote: Recognizing Silent and Weak Speech Based on Electromyography. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 595–604.

[35] Yann Soullard, Cyprien Ruffino, and Thierry Paquet. 2018.
CTCModel: Connectionist Temporal Classification in Keras.
(2018).

[36] STMicroelectronics. 2019.
iNEMO inertial module, 3D magnetometer, 3D accelerometer, 3D gyroscope, I2C, SPI.
(2019).
www.st.com/en/mems-and-sensors/lsm9ds1.html

[37] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018.
Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. ACM, New York, NY, USA, 581–593.
DOI:http://dx.doi.org/10.1145/3242587.3242599

[38] László Tóth, Gábor Gosztolya, Tamás Grósz, Alexandra Markó, and Tamás Csapó. 2018.
Multi-Task Learning of Speech Recognition and Speech Synthesis Parameters for Ultrasound-based Silent Speech Interfaces.
(09 2018).

[39] Michael Wand, Jan Koutník, and Jürgen Schmidhuber. 2016.
Lipreading with Long Short-Term Memory.
CoRR abs/1601.08188 (2016).
http://arxiv.org/abs/1601.08188