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

暦本 純一[1,2,a)]　木村 直紀[1,b)]　河野 通就[1,c)]

概要：
音声によって操作されるデジタル機器の利用可能性は急速に拡大している．しかし，音声インタフェースの使用状況は依然として制限されている．たとえば，公共の場で話すことは周囲の人に迷惑になり，秘密の情報を話すことができない．本研究では，超音波映像を用いて，利用者の無発声音声を検出するシステムを提案する．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の情報から，利用者が声帯を振動させずに発話した発声内容を認識する．超音波画像の系列から音響特徴を生成する２段階のニューラルネットモデルを提案する．提案モデルにより，合成したオーディオ信号が既存の無改造のスマートスピーカーを制御できることを確認した．これにより，人間とコンピュータが緊密に連携した種々のインタラクションが可能になり，新しいウェアラブルコンピュータが構成可能になる．また，咽頭の障害，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技術基盤につながるものである．

1. はじめに

音声対話によって制御可能なデジタル機器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，カーナビゲーションシステム，などの様々なデジタル機器は音声で制御可能になり，インタラクションの主要な手段としての音声対話の重要性が増している [41]． 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． 音声認識技術の進歩と音声合成の自然さにより，音声対話は人間とコンピュータとの対話において不可欠なプロセスとなっている． 利用者が運転，料理，家事，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，利用者は，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ながら，音声対話を用いて他の機器を操作することができる．

しかし，公共の場所での音声の使用には制限がある． 周囲の人に迷惑になることに加えて，個人情報や機密情報を発声することはできないし，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．

これらの課題を克服するために，サイレント音声認識の研究が進められている [6], [13], [45]．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，音声を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声のスキルを転用でき，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エア型のイヤフォンとこの技術を併用すると，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新しいウェアラブルコンピュータを構成することができる．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の，コミュニケーションを補助する技術としても利用できた場合には，高齢化社会に向けて有用な技術ということができる．

サイレント音声認識では，従来は映像方式がよく研究されてきた．話者の口唇画像または顔全体の映像をカメラによって撮影し，これらの画像から発話の内容を推定する [48]．しかし，映像方式では，顔の前方にカメラを設置する必要があり，その形状はウェアラブルまたはモバイル用途には不適切である． 「非可聴つぶやき」(Non-audible murmr, NAM) [18] では，利用者の喉に装着されたマイクロフォンで発話を認識しようという試みである．この場合，利用者は，声帯振動を伴わない呼吸音で話すが，マイクによって音を取得するためには，利用者のつぶやき音を

1 東京大学大学院情報学環・学際情報学府
2 ソニーコンピュータサイエンス研究所
a) rekimoto@acm.org
b) kimura.movie.edit@gmail.com
c) mchkono@acm.org

完全になくすことはできず，近くの他の人々に聞き取られてしまう可能性もある．筋電図 (EMG) によって口腔に近い筋肉の動きを推定することで，口頭発話を推定しようとする試みが行われている [28], [43]．現状では，EMG による自由発話の推定は困難であり，口腔運動から規定のコマンドを選択するジェスチャー認識の水準に留まっている．この場合，検出可能なコマンド数は限定されており，利用者は，筋電を駆動するための新たなジェスチャースキルを習得しなければならない．

上記のアプローチを取る代わりに，本研究では超音波イメージング [8] による無発声音声認識に着目する． 超音波イメージング技術は，体内に放射される超音波の反射時間を測定することにより，体内の内部状態を認識するもので，医療目的で体内の状態を把握するために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 (例えば Vscan Extend, General Electronic Company)．顎の下に小さな超音波イメージングヘッドを取り付け，超音波イメージングにより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であれば，実際に声帯を振動させずに話すことで音声対応装置と通信するための有用な機構となる．

超音波イメージングによるサイレントボイスインタラクションは，他のアプローチよりも二つの利点があると考えている． 第一に，超音波撮像センサーを小型化することができ，襟のように目立たない形状の装置を構成することができる． この構成は，ウェアラブルなサイレントボイスシステムを設計する上で重要となる． 第二に，口腔内の状況を認識することによって，外部から見ることができない舌の動きを測定し，より正確に音を再生できる可能性がある．

超音波イメージングを用いた無声音発話に関するこれまでの研究はいくつか存在しているが，それらの多くは，口唇や顔の画像の補助として用いるもので，利用者は顔の前にカメラを置かなければならない [21]． この構成は，ウェアラブルインタフェースデバイスとして使用される場合には制限がある．超音波イメージングとディープニューラルネットワークにより発話を認識しようとしている事例があるが，声帯を振動させて発話している場合の超音波映像を音声なしで認識する基礎実験に留まっている [5], [46]． それに対し，本研究では，超音波画像にのみ基づいた “SottoVoce” と呼ばれるサイレントボイスインタラクションシステムについて報告する．畳込みニューラルネットワークを利用し超音波エコー映像から音声を復元する機能を提供し，利用者が声帯を振動させずに発話の口の動きだけを再現した場合の音声復元の検証を行う．また，本方式の実用性を証明するために，既存のスマートスピーカー（Amazon Alexa) を無改造で用いて，復元音声により制御できることを示す．

2. 関連研究

2.1 サイレントスピーチ

サイレント音声インタフェースは，さまざまな技術と方法を使用して研究されている [6], [26]．音声情報を使用せずに，話題の発話を推定するには，口の画像 [45], [48] または顔全体の画像 [9] を使用する．Electromagnetic articulography(EMA) による手法が試されている [3], [49]．磁石を利用者に添付する手法 [11], [15], [19]，EEG [38] と EMG [27], [32], [33], [43] を利用する手法も研究されている．Kapur らが提案した最近の例では，内部音声の間に被験者の神経筋信号を感知するために複数の電極を使用している [28] ．上記で紹介した方法に加えて，顔面の電位による方法も試みられている [18], [42].

超音波イメージングによる無発声認識も試みられている [7], [22]． 唇の超音波とビデオに基づいて歌声母音を合成する [25]，舌と唇の超音波と光学画像を使用したサイレントスピーチインタフェース [21], [23]，超音波画像から舌の動きのアニメーションを自動的に生成するためのマッピング技法 [10]，などが研究されている．

ニューラルネットワークを利用したサイレントボイス認識の試みとしては，口唇画像の認識に LSTM (long short term memory) を用いるもの [48]，CNN を利用するもの [45]，超音波イメージングの解析にニューラルネットワークを用いるもの [5] がある．また，ニューラルネットワークにより F0（基本周波数）を推定するもの [17] がある．Eigentongues は複数画像の主成分を利用した固有顔 (Eigenfaces) の手法 [47] を舌形状の解析に適用している [20]．これらのアプローチは，現状では有声発声時の評価にとどまっており，無声発声時（声帯を振動させずに，発声時と同様に口を動かす）の検証はなされておらず，スマートスピーカーなどの既存の音声対話機器の制御の評価も行われていない．

1995 年に発表された Glove Talk II [12] は，ニューラルネットワークを用いた音声シンセサイザの 10 個の制御パラメータに手振りを変換するシステムである． しかし，システムを使用には熟練を要し，他者から聞き取れる水準の音声を発声させるために，ピアノ演奏の素養のある実験参加者の 100 時間を超えるトレーニングが必要だった．

2.2 関連する身体センシング手法

EarFieldSensing [34] は，電界感知に基づくジェスチャー認識技術である． CanalSense [2] は，顔が動いたときに起こる外耳道の気圧の変化を感知する． Tongue-in-Cheek [14] は，顔ジェスチャ認識のために X バンドドップラーレーダを使用して舌の動きを感知する． EMG [51] または脳と筋肉の信号感知の組合せを使用する他の方法 [37] も注目に値する．EchoFlex は，超音波イメージングを使用して前腕の筋肉の動きを認識する [35].

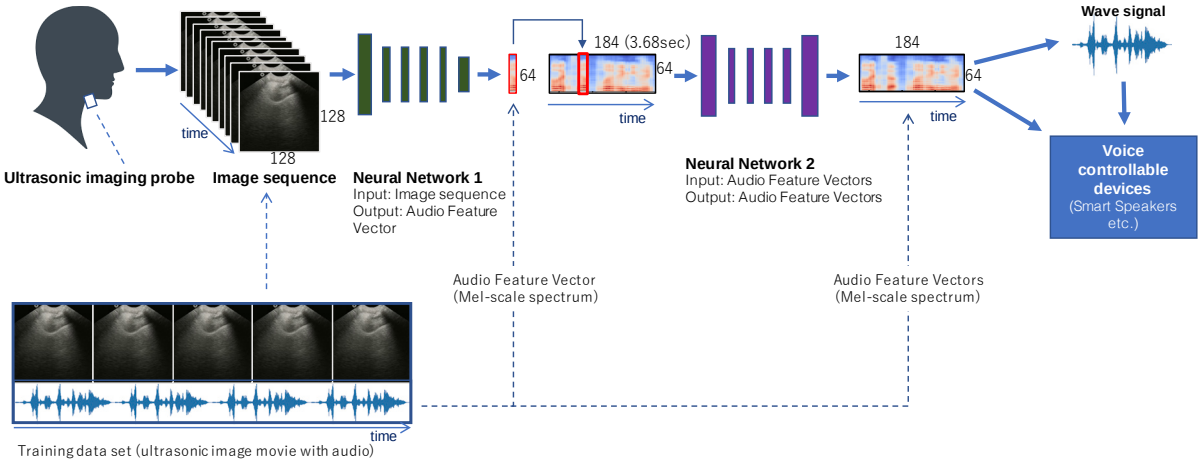

図 1　SottoVoce システム構成

2.3　知的エージェントとのコミュニケーション

モバイルデバイスやスマートデバイスの開発により，さまざまな手段によりデバイスと対話する機会が増えている [30]．入力方法として音声をつかうことができる [41]．このような音声制御の利用者インタフェースに関する研究がある [36]．音声で制御できる個人化エージェントやスマートスピーカと人間とのインタラクションの評価研究も多くなされている [31], [39], [44]．

3.　SottoVoce システム構成

提案システムの構成を，図 1 に示す．このシステムの目的は，ある時系列表現 (超音波画像の列) を他の時系列表現 (音声) に変換することである．この目的は，テキスト読み上げシステム [50]，声色変換システム [1]，画像によるリップリーディングシステム [9] などと関連が深い．これらのシステムを参考に，提案システムは 2 つのニューラルネットワークによる構成を提案する．

第 1 のニューラルネットワーク（図 1 の ‘Network 1’）は，超音波画像の列を音響特徴ベクトル（Mel スケールスペクトラム）に変換する．変換された音響特徴ベクトルは音声に復元可能な音響特徴ベクトルの列として接続される．そして音響特徴ベクトルの品質を改善するために，第 2 のニューラルネットワーク（図 1 の ‘Network 2’）を利用する．このネットワークは，音響特徴ベクトルの列を入力として，それを（より自然な）音響特徴ベクトルの列に変換する．

このニューラルネットで，利用者が有声発話した場合の口腔内の超音波映像と，対応する音声を用いて学習を行う．学習後，利用者が無声発話した場合の口腔内映像から音声が復元できるかを検証する．ニューラルネットは利用者依存で，学習時，実行時ともに同一の利用者が使用することを想定している．

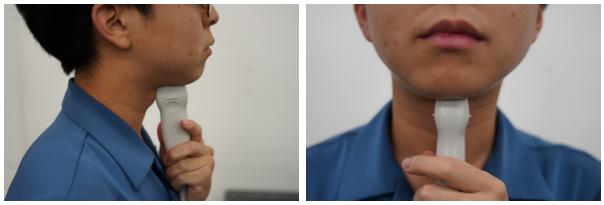

図 2　超音波イメージングプローブの装着位置

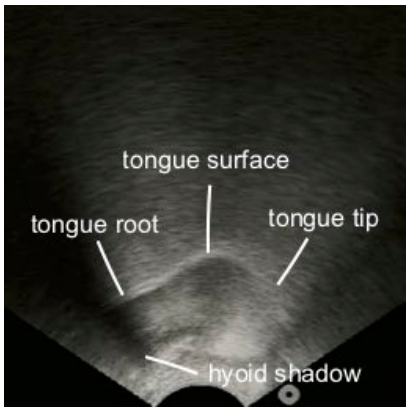

図 3　下顎に取り付けた超音波イメージングプローブにより取得された超音波エコー映像

3.1　超音波イメージング装置

超音波イメージング装置として CONTEC CMS600P2 Full-Digital B 超音波診断システムを使用した．利用者は $3.5MHz$ の凸型超音波イメージングプローブに計測用ゲルを塗布し顎の下に取り付けて使用する (図 2)．このシステムは，ディスプレイモニタに接続されるスクリーン出力ポートがある．このポートからの画像出力を画像デジタイジング装置により MPEG-4 画像ファイルに変換する．図 3 に得られた超音波画像の例を示す．

超音波画像システム内部での処理のため，音声キャプチャに対して超音波映像は遅延がある．超音波映像と音声を関係を調査し $300ms$ の遅延があることを確認した．トレーニングデータにはこの遅延を補正したものを用いる．

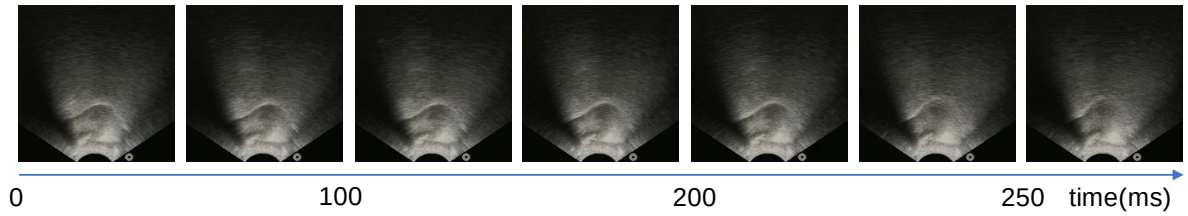

図 4 “Alexa” と発音する瞬間の超音波エコー映像の系列

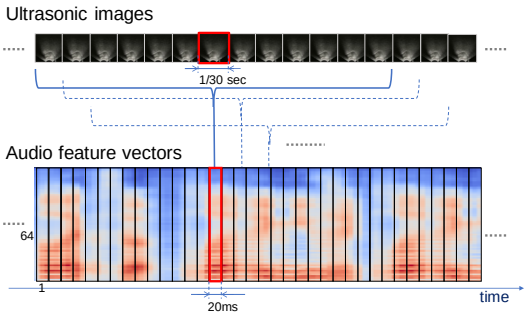

図 5 K 個の超音波映像の系列から，その中心位置の時刻に該当する音響特徴ベクトルを対応させ，学習データとする

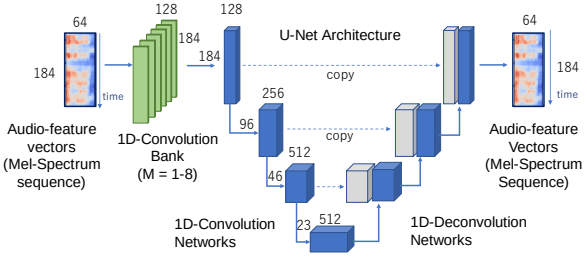

図 6 Network 2 による音響特徴ベクトル列の改善 (注：1D Convolution Bank の図示の整合性のため，入出力の音響特徴ベクトルの縦方向を時間軸として図示している)

3.2 Network 1

Network 1 は一連の K 個の超音波画像 (白黒，画素数は 128×128) を入力とし，n 次元の音響特徴ベクトル (Mel スケールスペクトラム) を出力として生成する． 現状，$K = 13$，$n = 64$ としている． 超音波画像のフレームレートは毎秒 30 フレームであるので，K 枚の超音波画像の持続時間は 400ms である． この時間は，発話の静的および動きの特徴を包含すると考えている． 超音波画像の例を図 4 に示す．舌の形状などが明確に取り出されているのがわかる．学習用には，有声発声時に採取した超音波映像と発話音声を用いて，画像系列と音声ベクトルの組を作っていく．音声は 1/50 秒ごとに Mel スケールスペクトラムに変換したものを用いている．画像と音声の採取レートが異なるが，K 枚の画像系列の中心となる画像の時刻に最も近い時刻の音響ベクトルを対応づけている．この作業を画像系列を順次ずらしながら繰り返し適用し，学習用のデータセットを作成する (図 5)．認識時には、1/50 秒ごとにその時刻に対応する画像フレームを求め、それを含む前後 K 枚の超音波画像から音響特徴ベクトルを求めていく．

K 個の超音波映像シーケンスの中心の時間位置に対応する音響特徴ベクトルが音波データから抽出され，Network 1 がそれを生成するように訓練する．

Network 1 は，畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) に基づいている．具体的には 4 つの層，*Conv2D - LeakyReLU - Dropout - BatchNormalization* とそれに続く 6 つの層 *Flatten - Dense - LeakyReLU - Dropout - Dense - LeakyReLU* で構成されている． Network 1 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 [29] を使用した．

3.3 Network 2

音質を向上させるために，Network 2 は音響特徴ベクトル列を入力とし，その入力と同じ長さの音響特徴ベクトル列を生成する．Network 1 で生成された音声の品質が部分的に不明瞭であった場合（たとえば “Alexa, play *sic” のように*の部分が不明瞭だった場合) に，それを改善する (“Alexa, play music.”) ことを期待している．

ニューラルモデルは，1 から M(現状は $M = 8$) のカーネルサイズを持つ 1 次元畳み込みフィルタ (it Conv1D) の組と，U-Network [40] を利用した Encode Decode ニューラルネット (図 6 *Conv1D - MaxPooling(strides = 2) - LeakyReLU - Dropout*) を組み合わせている． 1D 畳み込みバンクは，入力シーケンスの時間軸的な広がりの狭い範囲から広い範囲までのコンテキストをモデル化することを想定している． 次の U-Network は，Encode-Decode ネットワークにより生成する出力ベクトルの品質を向上させることを期待している． 最終出力として，Network 2 は入力と同じ時間長さの Mel スケールスペクトラムベクトルを生成する．

Network 2 を訓練するために，Network 1 は，訓練用超音波ビデオクリップの画像からの入力として Mel スケールのスペクトラムベクトルを作成し，出力と同じトレーニングビデオクリップのオーディオから同じ長さの Mel スケールスペクトラムベクトルを作成するために使用される． Network 1 の場合と同様に，平均二乗誤差を損失関数に，Adam を最適化関数として使用した．

簡単のため，入力と出力の時間間隔は同じ値に固定され

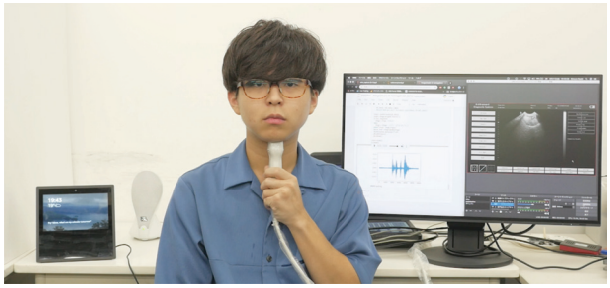

図 7 学習・再生実験環境

ている (現在，3.68 秒が使用されている)． この時間は，多くの典型的な音声コマンドを包含する長さである．音声コマンドがこの長さに満たない場合は無音声でパディングされる．

3.4 音声合成

ニューラルネットワーク処理に続いて，Mel スケールスペクトラムの音響特徴ベクトル列を，Griffin Lim 手法 [16] により音声波形に復元する． この変換は，Network 1 の出力からも，Network 2 の出力からも可能である．

生成された音波はオーディオスピーカから再生され，スマートスピーカなどの近くのサウンド制御デバイスを無改造で制御することができる．これにより，種々の音声制御可能な情報機器を制御できる可能性が示された．また，生成される音声は，利用者が，自分の口腔の動きがどのような音声を生成するのかを知る手段としても重要である．

音声波形信号を音声制御可能な装置の音声入力情報として実際に再生せずにそのまま取り込むことも可能である．その場合は，Bluetooth, 光通信などの手段により，利用者の音声出力を情報機器に伝達することが可能である．また，音声波形まで変換せずに，音響特徴ベクトルを，情報機器の音声認識の入力として利用することも可能である．

3.5 学習

訓練データの作成に関しては，2 人の実験参加者 (男性 28 人，男性 24 人) が顎の下に超音波プローブを取り付け，様々なスピーチコマンドを発声してもらい，超音波映像と音声を採取した．各実験参加者ごとに約 500 の音声コマンドを収集した（図 7）．具体的に用いられたスピーチコマンドは “Alexa, play music,” “Alexa, what’s the weather like,” “Alexa, what time is it,” ”Alexa, play jazz”などである．

記録されたデータは Network 1 を訓練するためにまず使用する． 超音波画像は 128×128 にスケーリングされ，ニューラルネットの入力側の訓練データとして使用される． 同時に記録した発話音声は Mel スケールスペクトラムに変換し，出力側の訓練データとして使用された．

訓練した Network 1 を利用して，Network 2 の学習を行う．Network 1 を利用して，音声コマンドに相当する音響特徴ベクトルの列を生成する．これをニューラルネットの学習側のデータとして用い，実験参加者が発話して得られた音声データから生成した音響特徴ベクトルの列をニューラルネットの出力側の学習データとして用いる．これにより，Network 1 が生成した音声の品質をより大域的に改善することができる．

Network 2 のテストセットの数は，採取した音声コマンド発話数 (約 500) と同じである． テストセットの数を増やすために，data augumentation の手法を用いた．具体的には，入力側の音響特徴ベクトルの値を正規分布乱数で増減させたデータセットによりテストデータ数を増大させた．

我々のモデルは話者に依存するので，Network 1 と Network 2 は両方とも各話者に対して訓練される． Network 1 が最初に訓練され，Network 2 の訓練のためのデータセットを作成するために使用される．

3.6 実装

上記のネットワークモデルを，深層学習ライブラリ Keras [4] を用いて実装した． 実行には Ubuntu 16.04, Intel Core i7-8700K CPU (3.7GHz)，GPU ボードとして，NVIDIA GeForce GTX 1080TI を用いた．Network 1 の学習に約 4 時間を要した．Network 2 の学習は 1 時間以内に完了した．

超音波イメージング装置を，ニューラルネットワークを実行するコンピュータ (Ubuntu マシン) に直接接続することができなかったため，単純なサーバークライアントプログラムを開発した． 超音波撮像装置を制御するコンピュータが，画像の系列をネットワーク経由でニューラルネットワーク実行システムに送信する．生成された音声データもネットワーク経由で返信され，再生することができる．3.68 秒の音声コマンドを超音波イメージング装置で採取し，ネットワーク経由で転送し，ニューラルネットワークで処理をした後音声データをネットワーク経由で返信するまでの end-to-end の処理時間は 2.36 秒だった． ネットワーク処理を含まない総処理時間 (ビデオ処理，ニューラルネットワーク処理，Mel スケールスペクトラムのオーディオ波への変換を含む) は 2.61 秒 だった．

4. 結果

超音波画像を音に変換した結果を図 8 に示す．図中，上の段のグラフは音響特徴ベクトル (Mel スケールスペクトラム) を示し，下の段のグラフは対応する波形を示す．Net1 とラベル付けされたグラフは Network 1 の結果であり，Net2 は Network 1 + Network 2 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 Mel 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る．したがって，最後のものは学習における ground truth とみなすことができる．

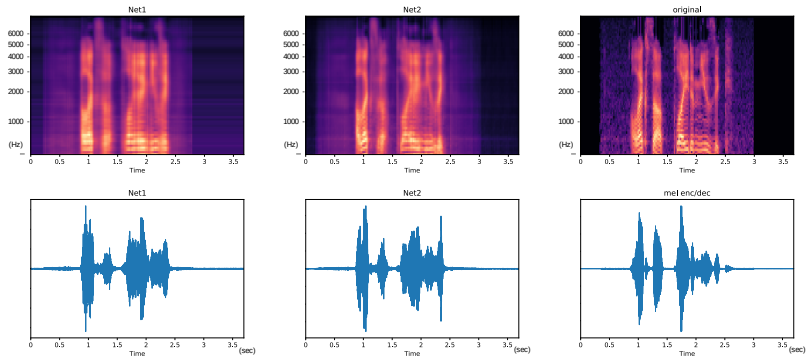

図 8 結果: Net1: Network 1 の出力; Net2: Network 2 の出力; original: 元の学習データの音声を音響特徴ベクトル系列に変換したもの (学習時の 'ground-truth' となる)

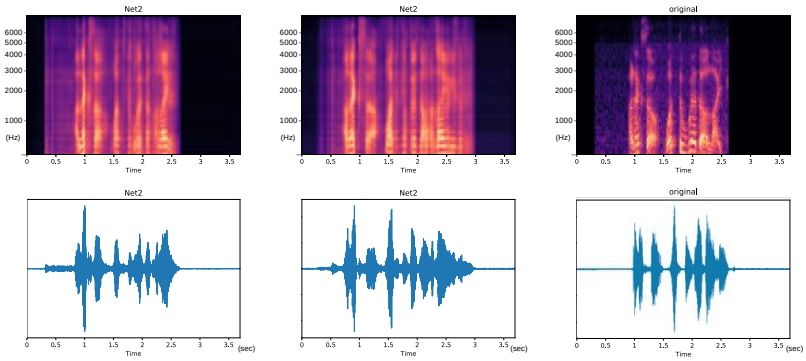

図 9 生成音声の比較 (左: 有声発話時に取得した超音波映像から生成した音声 (Network 2 出力), 中: 無音声発声時に取得した超音波映像から生成した音声 (Network 2 出力), 右: 有声発声時に取得した音声から生成した音響特徴ベクトル (学習時の'ground-truth'))

生成された音声は，既存の (変更されていない) スマートスピーカー (Amazon Echo および Amazon Echo Show) でテストされ，生成されたサウンドがスマートスピーカーを制御できることが確認された．

表 1 に，Network 1，Network 1 + Network 2，および元の（Mel スケールが符号化され，デコードされた）認識率を示す．図 8 や，聴感上も，Network 1 と Network 2 の出力の違いは必ずしも明確ではないが，Network 1 と Network 2 の組み合わせが認識率を改善することを確認した．また，トリガーワード（「Alexa」）は常に非常に明確に再生成されていることがわかった．これは，その単語が単にトレーニングセットの中で最も発音された単語であったことが可能性として考えられる．改善を定量的に評価するため，Google 社の Cloud Speech-to-Text engine [24] を用いて WER(Word Error Rate) の測定を行った．このときの WER は GT，Network 1，Network 2 それぞれ 20.61%，41.03%，33.56%であり，このことからも Network 2 で生成された音声の品質は向上しているといえる．

5. End-to-End 認識での評価と知見

実際の end-to-end の音声から音声への変換について検証を行った．この場合，利用者は実際に音を出さずに口頭の動きのみで音声コマンドを発話するように指示され，口

表 1 音声認識成功率（既存のスマートスピーカーを無改造で用いた場合)．"GT" は有性発話の音声を音響特徴ベクトル列に変換し，また音声データに複合した音声を利用した場合の認識率で，学習時の 'ground-truth' となる

	user A	user B	total
Network 1	60.0%	25.0%	42.5%
Network 1 + Network 2	65.0%	65.0%	65.0%
GT	90.0%	90.0%	90.0%

腔の動きは超音波プローブによって記録される．得られた画像シーケンスは，提案されたシステムによって音声に変換された．図 9 は，利用者が声を出しているときと，利用者が声を出していないときの生成音の比較を示している．ここでいう音声を発声させない発話とは，利用者に可能な限り静かに発話することを指示したものであって，息を止めることを指示していたわけではない．したがって，僅かな音の漏れは確認された．この際の音声レベルを音圧計で測定したところ，その平均は 37.14dB(A) であった（合計 20 個 Alexa コマンドの発話による測定の平均値)．この際周囲の騒音レベルは 31.0dB(A) であった．

利用者が声を出さない場合の口腔内の動きと，実際に声を出したときの動きはには微妙な違いがあり，音声なしの画像によって生成される音質は，音声付きの画像によって生成される音質ほど良くはなかった．(図 9)．一方，利用

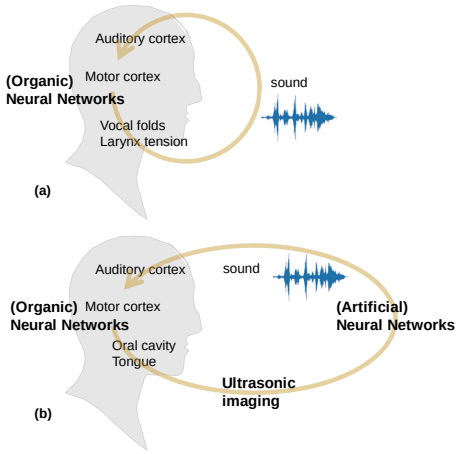

図 10 Human-AI integration: 人工のニューラルネットと人間が緊密なフィードバックループを形成することで，人間側の能力も向上する

図 11 本研究を利用したウェアラブルコンピュータの想定図．咽頭に小型の超音波プローブを装着し，骨伝導または穴あきのイヤフォンを装着し，音を外部に漏らさずに音声エージェントと対話を行うことが可能になる．

者は，音声なしの画像によって生成された音声を聴くこともできるので，口腔の動きを変化させて，音質がより良くなるように変更しようする傾向が観察された．この場合，利用者は自分の声ではなく，ニューラルネットによって生成した音声を聞くことで，自分の無発声発話能力を向上させようとしていると考えられる．

6. 議論

6.1 Human-AI integration

前節での観察は，人間と AI との間に興味深い関係が存在することを示唆している． AI を自律的または分離した存在とみなすのではなく，AI を人間の一部として考えることができるかもしれない． その場合，(人工) ニューラルネットワークの初期性能が完璧でない場合であっても，利用者は徐々にその性能を学び，改善することができる．

これは人々が基本的なスキルを学ぶ方法と似ている． 通常，発声を学習するときには，口腔，舌，声帯を駆動する運動皮質と聴覚皮質との協調は，より良い会話性能を得るためのフィードバックループを形成する (図 10 (a)) ． このループを拡張することによって，われわれの（オーガニックな）ニューラルネットワーク (我々の脳) と，人工ニューラルネットワークの総体がフィードバックループを形成することになる (図 10 (b)). これは，この形成を「人間と AI の統合 (Human-AI Integration)」と名づけられるのではないかと考えている．従来のインタラクティブシステムでも、このように利用者がシステムからのフィードバックを通じて自らの能力を改善することはあったが、そのループにニューラルネットワークが介在することで、どちらか片方では得られない効果を生み出すことができるかもしれない．

この点，1995 年の Glove Talk II の研究は，人間と AI の統合に関する先駆的研究の 1 つとして興味深い [12]．この研究では，利用者は手のジェスチャーによるボイスシンセサイザーの制御を習得する．３つの単純なニューラルネットワークが使用され，実験参加者（ピアニストであった）は可聴音を生成するために 100 時間以上を要した．より良いニューラルネットワークと学習者の組み合わせは，この学習時間を短縮することができると考える．

6.2 逐次的な音声変換

我々の現在の設計では，得られた超音波画像シーケンスは，音声コマンドの粒度 (約 3.68 秒) で音声に変換される． これは，Network 2 が固定長の音声表現シーケンスをとるためである． しかし，利用者の練習の観察に基づいて，音を段階的に生成する方が良いはずであるため，利用者はより良い声を生成するために口腔運動を学習するためのより緊密なフィードバックループを有することができる．Network 1 においても，現状では音響ベクトルの発生時刻の前後を含む K 枚の画像系列を利用しているが，音響ベクトルの発生時刻およびそれ以前の時刻の画像系列のみを用いた方法も検証中である．これにより，少なくとも Netowrk 1 においては遅延なく音声を生成することが可能になると期待している．

6.3 発声困難者に対するインタフェース

声帯の損傷を受けた人々も私たちの研究を利用することを期待している． Human-AI integration の章で説明したように，人は，声帯が機能しないにもかかわらず，口や舌を正しく制御して音を生成する方法を学ぶことができる．

今回，口腔の超音波映像のみからイントネーションが復元できたが，この知見は電動式人工咽頭への応用可能性を提供する．既存の人工咽頭は規定の周波数しか発声することができず，発声する音声が「ロボットボイス」のように人工的に聞こえてしまう．本研究の方式を応用して，より自然なイントネーションが復元できるかもしれない．

6.4 身体への連続的な超音波の放射

超音波が体内に持続的に放出されるときの人間の臓器へ

の影響は不明であるが，一般的に考えられる影響は発熱作用である．一方で，超音波画像装置は医療の現場で一般的に使用されており，その技術自体は十分に安全であるといえる．唯一この手法が利用できない場合として，眼球への照射が挙げられる．本研究では，口腔内を超音波画像で撮影しようとする性質上，口腔内の空気層によって超音波の伝達が妨げられることもあり，眼球や他の身体部位への暴露はないと考えられる．したがって，超音波画像装置を本研究で利用することは十分に安全であり，著者らの所属機関の倫理審査委員会でも実験について承認されている．

また，超音波を常時放出する必要はなく，単純なトリガ機構を組み合わせて，超音波の放出を開始および停止することができると考えている．例えば，加速度センシングにより，実際に音声を発することなく，（無音の）音声コマンドを開始するための顎の動きを検出することができる．

6.5 超音波プローブの設置位置

超音波プローブは同じ位置・角度に当てることで明瞭な音声が生成できると予想されたため，なるべく同じ位置に当てるよう実験参加者に指示したが，実際にはデータセット収集時・テスト時両方の場面においてわずかな角度・位置ずれは発生していた．今回はわずかなずれがあってもスマートスピーカーを操作可能な音声を生成できたが，センサを当てる位置を厳密に定めることにより，より明瞭な音声を安定して生成できる可能性がある．また現状ではゲルを塗布したプローブを用いているが，センサー部を皮膚に固定する際に半固形ゲルを使用するなどを検討したい．常に正確に同じ位置に装着可能な，専用装着器具としての開発は，ウェアラブル用途での利用可能性検証のみならず，生成の精度向上においても重要であると考える．

また，顔の正面にセンサを設置しなくて良い点は超音波エコーを用いる利点であり，下顎から照射することで最も多くの情報を捉えることができると考えたため，その他の箇所での実験は行っていないが，より側面に近い位置や角度を変えることで異なる結果を得る可能性がある．最も最適な位置を探ることは今後の研究対象である．

6.6 他のモダリティとの融合

最後に，この研究は他の様式を排除することを意図したものではないことを指摘する． EMG，加速度計，およびNAM マイクロフォン等の情報を組み合わせることで，音声認識の質が向上する可能性がある． これらの様式の組み合わせを調査することは，将来の研究の対象である．

7. 結論

本稿では，超音波画像を用いた音声対話の方法を提案した．2 つのニューラルネットワークを使用して，音声なしの利用者の擬似「発話」を，スマートスピーカなどの既存の音声制御可能なデバイスを操作するために使用できる音声に変換する．

この結果に続いて，ウェアラブルコンピュータの今後のフォームファクタは，骨伝導性イヤホンまたはオープンエアイヤホンを備えたカラータイプの超音波プローブと組み合わせることを想定している（図 11）．この構成では，利用者は，音声を発することなく常に音声制御可能なアシスタントを呼び出して応答を得ることができる．

謝辞 貴重な助言を頂いた査読者・プログラム委員に深謝いたします．

参考文献

[1] Dabi Ahn. 2017.
Voice Conversion with Non-Parallel Data.
https://github.com/andabi/deep-voice-conversion.
(2017).

[2] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017.
CanalSense: Face-Related Movement Recognition System Based on Sensing Air Pressure in Ear Canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 679–689.
DOI:http://dx.doi.org/10.1145/3126594.3126649

[3] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux, and Blaise Yvert. 2016.
Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces.
PLOS Computational Biology 12, 11 (11 2016), 1–28.
DOI:http://dx.doi.org/10.1371/journal.pcbi.1005119

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

[5] 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*.

[6] 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

[7] B. Denby and M. Stone. 2004.
Speech synthesis from real time ultrasound images of the tongue. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. I–685.
DOI:http://dx.doi.org/10.1109/ICASSP.2004.1326078

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

[9] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. 2017.
Improved Speech Reconstruction from Silent Video.
ICCV 2017 Workshop on Computer Vision for Audio-Visual Media (2017).

[10] Diandra Fabre, Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda, and Pierre Badin. 2017.
Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract.
Speech Communication 93 (2017), 63 – 75.

[11] M.J. Fagan, S.R. Ell, J.M. Gilbert, E. Sarrazin, and P.M.

Chapman. 2008.
Development of a (silent) speech recognition system for patients following laryngectomy.
Medical Engineering & Physics 30, 4 (2008), 419 – 425.

[12] Sidney Fels and Geoffrey Hinton. 1995.
Glove-TalkII: An Adaptive Gesture-to-formant Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 456–463.
DOI:http://dx.doi.org/10.1145/223904.223966

[13] 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

[14] Mayank Goel, Chen Zhao, Ruth Vinisha, and Shwetak N. Patel. 2015.
Tongue-in-Cheek: Using Wireless Signals to Enable Non-Intrusive and Flexible Facial Gestures Detection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 255–258.
DOI:http://dx.doi.org/10.1145/2702123.2702591

[15] Jose A. Gonzalez, Lam A. Cheah, James M. Gilbert, Jie Bai, Stephen R. Ell, Phil D. Green, and Roger K. Moore. 2016.
A Silent Speech System Based on Permanent Magnet Articulography and Direct Synthesis.
Comput. Speech Lang. 39, C (Sept. 2016), 67–87.
DOI:http://dx.doi.org/10.1016/j.csl.2016.02.002

[16] 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

[17] Tamás Grósz, Gábor Gosztolya, Lászlo Tóth, Tamás Gábor Csapó, and Alexandra Markó. 2018.
F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), 291–295.

[18] 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.
Silent Speech Interfaces.

[19] Robin Hofe, Stephen R. Ell, Michael J. Fagan, James M. Gilbert, Phil D. Green, Roger K. Moore, and Sergey I. Rybchenko. 2013.
Small-vocabulary Speech Recognition Using a Silent Speech Interface Based on Magnetic Sensing.
Speech Commun. 55, 1 (Jan. 2013), 22–32.
DOI:http://dx.doi.org/10.1016/j.specom.2012.02.001

[20] T. Hueber, G. Aversano, G. Cholle, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone. 2007.
Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Vol. 1. I–1245–I–1248.
DOI:http://dx.doi.org/10.1109/ICASSP.2007.366140

[21] 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

[22] Thomas Hueber, Elie-Laurent Benaroya, Bruce Denby, and Gérard Chollet. 2011.
Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface. In *INTERSPEECH*.

[23] Thomas Hueber, Gerard Chollet, Bruce Denby, and M Stone. 2008.
Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. In *Proceedings of International Seminar on Speech Production*. 365–369.

[24] Google Inc.
Clund Speech-to-Text.
https://cloud.google.com/speech-to-text/. (????).

[25] Aurore Jaumard-Hakoun, Kele Xu, Clémence Leboullenger, Pierre Roussel-Ragot, and Bruce Denby. 2016.
An Articulatory-Based Singing Voice Synthesis Using Tongue and Lips Imaging. In *ISCA Interspeech 2016*, Vol. 2016. San Francisco, United States, 1467 – 1471.
DOI:http://dx.doi.org/10.21437/Interspeech.2016-385

[26] Yan Ji, Licheng Liu, Hongcui Wang, Zhilei Liu, Zhibin Niu, and Bruce Denby. 2018.
Updating the Silent Speech Challenge Benchmark with Deep Learning.
Speech Commun. 98, C (April 2018), 42–50.
DOI:http://dx.doi.org/10.1016/j.specom.2018.02.002

[27] Chuck Jorgensen and Kim Binsted. 2005.
Web Browser Control Using EMG Based Sub Vocal Speech Recognition. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences - Volume 09 (HICSS '05)*. IEEE Computer Society, Washington, DC, USA, 294.3–.
DOI:http://dx.doi.org/10.1109/HICSS.2005.683

[28] 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

[29] Diederik P. Kingma and Jimmy Ba. 2014.
Adam: A Method for Stochastic Optimization.
CoRR abs/1412.6980 (2014).
http://arxiv.org/abs/1412.6980

[30] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017.
The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, New York, NY, USA, 555–565.
DOI:http://dx.doi.org/10.1145/3064663.3064672

[31] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018.
Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 881–894.
DOI:http://dx.doi.org/10.1145/3196709.3196784

[32] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. 2005.
Session independent non-audible speech recognition

using surface electromyography. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* 331–336.
DOI:http://dx.doi.org/10.1109/ASRU.2005.1566521

[33] Hiroyuki Manabe, Akira Hiraiwa, and Toshiaki Sugimura. 2003.
"Unvoiced Speech Recognition Using EMG - Mime Speech Recognition". In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*. ACM, New York, NY, USA, 794–795.
DOI:http://dx.doi.org/10.1145/765891.765996

[34] 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

[35] Jess McIntosh, Asier Marzo, Mike Fraser, and Carol Phillips. 2017.
EchoFlex: Hand Gesture Recognition Using Ultrasound Imaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1923–1934.
DOI:http://dx.doi.org/10.1145/3025453.3025807

[36] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018.
Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 6, 7 pages.
DOI:http://dx.doi.org/10.1145/3173574.3173580

[37] Phuc Nguyen, Nam Bui, Anh Nguyen, Hoang Truong, Abhijit Suresh, Matt Whitlock, Duy Pham, Thang Dinh, and Tam Vu. 2018.
TYTH-Typing On Your Teeth: Tongue-Teeth Localization for Human-Computer Interface. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '18)*. ACM, New York, NY, USA, 269–282.
DOI:http://dx.doi.org/10.1145/3210240.3210322

[38] Anne Porbadnigk, Marek Wester, Jan Calliess, and Tanja Schultz. 2009.
EEG-based Speech Recognition - Impact of Temporal Effects. In *BIOSIGNALS*.

[39] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017.
"Alexa is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 2853–2859.
DOI:http://dx.doi.org/10.1145/3027063.3053246

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015.
U-Net: Convolutional Networks for Biomedical Image Segmentation.
CoRR abs/1505.04597 (2015).
http://arxiv.org/abs/1505.04597

[41] 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

[42] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014.
The Tongue and Ear Interface: A Wearable System for Silent Speech Recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers (ISWC '14)*. ACM, New York, NY, USA, 47–54.
DOI:http://dx.doi.org/10.1145/2634317.2634322

[43] 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.

[44] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018.
"Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 857–868.
DOI:http://dx.doi.org/10.1145/3196709.3196772

[45] 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

[46] 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. In *Proc. Interspeech 2028*. 3172–3176.

[47] M. Turk and A. Pentland. 1991.
Face recognition using eigenfaces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 586591.

[48] 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

[49] Jun Wang, Ashok Samal, and Jordan Green. 2014.
Preliminary Test of a Real-Time, Interactive Silent Speech Interface Based on Electromagnetic Articulograph. In *Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*. 38–45.
DOI:http://dx.doi.org/10.3115/v1/W14-1906

[50] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017.
Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model.
CoRR abs/1703.10135 (2017).

[51] Qiao Zhang, Shyamnath Gollakota, Ben Taskar, and Raj P.N. Rao. 2014.
Non-intrusive Tongue Machine Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2555–2558.
DOI:http://dx.doi.org/10.1145/2556288.2556981