SilentMask: 口周辺の運動計測によるマスク型サイレントスピーチインタラクション

平城 裕隆[†1,a)] 暦本 純一[†1,†2,b)]

概要：サイレントスピーチインタラクション (SSI) は，有声ではない発声による音声インタラクションであり，スマートフォンなどの音声認識デバイスへの入力手段として，また発声困難者への支援として用いられる．従来から口の周辺を利用する SSI としてリップリーディングや筋電，超音波エコー，口蓋内の静電測位等を用いた手法が提案されてきたが，片手が塞がることや目立ちやすいといった課題があった．本研究では，マスクに取り付けた加速度と角速度のセンサを利用し，口周辺の運動を計測することで無声での発話を認識するマスク型の SSI を提案する．2 つの加速度センサを用いて 12 次元の口周辺の運動情報を取得し，深層学習を用いて解析したところ，21 種類の音声コマンドと発話していない状態の計 22 状態を 79.1%の精度で識別でき，表情と動作の計 6 種類に関して分類し 84.7%の精度で認識できた．また，主観調査を行い，加速度センサを直接皮膚に貼る手法に比べて長時間着用できるという結果を得た．本研究は，カメラ画像を用いないため照明条件によらず，目立たず片手を塞がないインターフェースであり，マスクの新たな可能性を提示するものである．

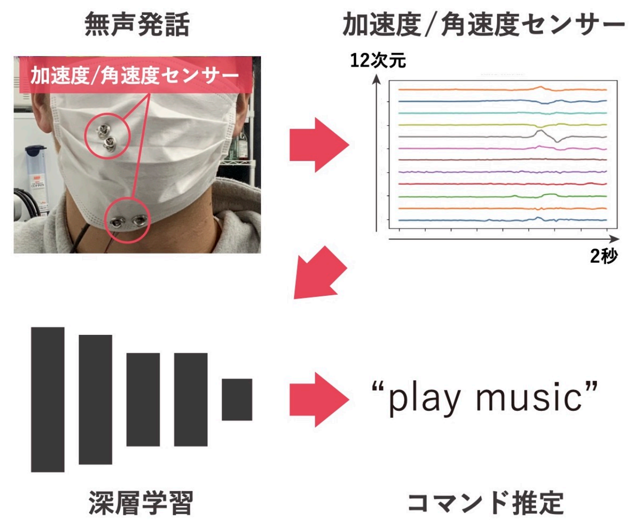

図 1 SilentMask は布製のマスクにセンサを固定し，周辺の皮膚運動計測によって無声発話を認識するサイレントスピーチインタラクションである．

1. 背景

電子デバイスと人間の対話環境は，音声認識技術や自然言語処理の発展によって身近な存在になってきた．特にスマートフォンやスマートスピーカーなどの音声検索や対話的な入力は，GUI など他のインターフェースと同様に日常的に用いられている．

他方で，発声型の音声インターフェースは公共空間での発声が憚れることや，騒音・他者の声など意図しない音声が入力され得る環境においては利用されにくいといった課題がある．また，発話内容のプライバシーの点でも個人情報の入力など他者に聞かれるべきでない情報を伝達するのには適していない．

これらの課題を解決するためにサイレントスピーチインタラクション (SSI) が研究されている [1]．これは声帯の振動を伴わない発声において生じる身体の変化をデバイスで計測することによって発声を推定し，インタラクションとして利用するものである．SSI は有声発声インタラクションの代替手段としてだけでなく，声帯振動での発声が困難な者への補助手段としても用いられている．

SSI は様々なセンサーや測定位置が提案されている．特に口周辺に着目した手法としては口腔内の静電センサー [2] や筋電による口周辺の運動計測 [3]，などがある．一方で，それらはデバイスが目立ち不自然なものや，照明条件など外部の環境によって計測に影響が生じるもの，片手の補助を伴い音声インターフェースとしての自由度を低下させているものなど各々課題がある．

そこで本研究ではマスク型デバイスによる SSI である

†1 現在，東京大学大学院情報学環
†2 現在，ソニーコンピュータサイエンス研究所
a) hiraki-uts1@g.ecc.u-tokyo.ac.jp
b) rekimoto@acm.org

SilentMask を提案する．これは市販のマスクに加速度・角速度センサーを取り付けて無声発話時の皮膚の運動情報を取得し，深層学習を用いて発話や動作を推定する．SilentMask はマスクという衛生用途として日常的に用いられるデバイスを拡張することで，不自然に目立つことなく，観測する環境に依らず，片手の補助を必要としないインターフェースを実現している．

本研究ではスマートスピーカーへの入力として用いられる 21 の音声コマンドと 1 つのを用いて行い，79.1%の認識率を示した．また，表情と動作の計 6 種類に関して分類し 84.7%の精度で認識できた．主観評価も加えて行い，皮膚にセンサーを貼り付ける手法と比較して長時間着用可能であることを示した．

2. 関連研究

2.1 サイレントスピーチインタラクション

SSI では身体の測定する箇所に応じたセンサ構成が提案されており，一概に分類するのは難しい．Kapur ら [3] は身体に対して侵襲的か否かで大別している．暦本ら [4] は画像方式のリップリーディング，非可聴つぶやき，超音波画像，筋電図，口の中または周辺のデバイスの 5 つに分類し各手法の課題を纏めている．本研究の立ち位置は「侵襲的でなく」かつ「口の中または周辺のデバイス」に該当し，加えて皮膚運動の計測による無声発話の推定を行うものである．

マスクのように非侵襲なデバイスによるサイレントスピーチは，様々なセンサによって行われており，RGB 画像によるリップリーディング [5][6]，超音波エコー [7][8]，マイク [9]，筋電図 (EMG)[3][10][11]，加速度センサ [4] などが挙げられる．

口に直接触れる点では，福本 [9] はマイクを口の前面の非常に近い距離に設置し，吸気での発話を取得することで周囲に音が漏れない手法を提案している．マイクは片手で押さえる設計となっており音声インターフェースの自由度を低減させている，また吸気による音声入力は新たな発声の訓練を必要とする．

皮膚の運動計測に関しては，暦本ら [4] は加速度・角速度センサーを用いて顎下の皮膚運動計測を用いる手法を提案しており，22 の発声コマンドを 94%以上の精度で認識している．

2.2 マスク型デバイスによるインタラクション

顔の一部，特に口周辺を覆うマスクは衛生対策として日常的に用いられると共に，医療用途でも利用されている．マスクは顔の一部を覆うことができることから，表情の取得や口の開閉の判別などが研究されている．Mose ら [12] は，フォトリフレクタを用いて顔の運動を計測してパペットを操縦できるお面状のマスクを提案している．石井ら [13] はマスクの前面にディスプレイを設置し音声入力から動的に表情を変えてマスク越しに伝達する手法を提案している．山崎ら [14] はマスクの表面に格子状のタッチセンサを配置し口の形状を認識させている．Nam ら [15] は加速度センサー，LED の映像，マスク内の小型カメラの 3 つの手法で唇の運動を取得し表情の伝達を行っている．Lee ら [16] は反射型フォトリフレクタを用いて 12 の唇の運動と 11 の表情を分類している．

3. SilentMask

サイレントスピーチにおいてマスクは顔の表面と触れる面積が大きく，無声発話によって生じる口腔の運動を捉えることができると考えられる．本研究ではマスク型デバイスに加速度センサーと角速度センサーを取り付けて皮膚の運動を計測し無声発話を認識する手法を提案する．加速度・角速度センサーは小型で目立ちにくく，安価かつ省電力であるため，日常的に用いられるデバイスであるマスクのインタラクションに向いていると言える．

使用するセンサーは InvenSense 製の MPU-6050 であり，2 つのセンサーからなる 12 次元のデータを raspberrypi zero を用いて 60fps で取得する．センサーは市販の不織布のマスクにサージカルテープとピンバッチで固定している．取り付ける場所は図 2 の位置でありこれはマスクを装着した際に顎と右頬の位置に該当する．この位置は筋電センサーを用いた SSI で利用されており [3]，無声発話時の運動情報を含んでいると考えられる．

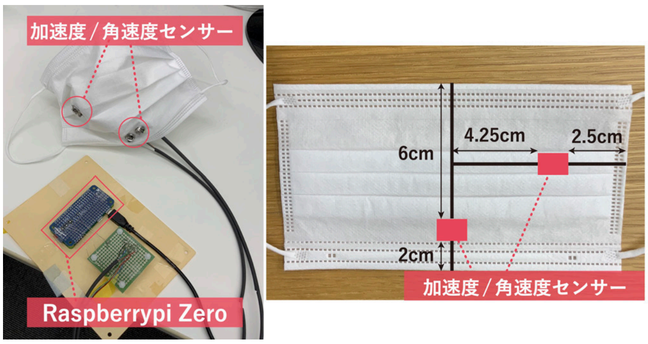

図 2　システムの構成

4. データの収集

4.1 音声コマンドの選択

サイレントスピーチインターフェースはどのコマンドを認識すべきかが確立されておらず，AlterEgo[3] では 0-9 の 10 単語，Lip-Interact[5] では頻繁に使われる 44 のコマンド，Derma[4] では 21 単語とまばらである．

今回は，単純で Alexa の入力にも用いられる 21 単語を用いた (表 1)．また，無声の発声を行っていない状態のセンサーデータを収集し，計 22 種類の発話状態に分離する．

表 1　音声コマンド

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

4.2　データの収集

データの収集は，日常的に有声発話を行っている 4 名 (男性 3 名・女性 1 名) で行った．各ユーザーは 22 個の各コマンドについて 10 回の発話を 1 セクションとして 4 セクション行った．

セクションごとにマスクを外して再度着用し，またセクションごとに有声発話と無声発話を交互に切り替えた．この 4 回のセクションを 1 セットとし，提案手法の場合と，皮膚に直接センサを貼り付けた場合の 2 セットをユーザーごとに行った．

発話の際は発声すべき単語を表示し，その後約 2 秒間の間のセンサーの値を記録した．ユーザー 1 名につき，10 回 × 4 セクション× 22 コマンドの計 880 回分のデータを取得した．

また，図 3 に示す 3 つの表情と 3 つの動作に関して同じく 40 回ずつ収集し，ユーザーあたり 240 回分のデータを取得した．

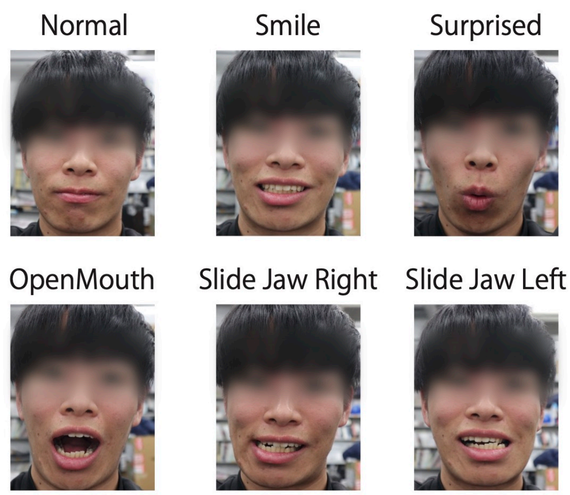

図 3　表情・動作

4.3　データの前処理

センサーの平均 fps は 60Hz であり，1 回の発話につき 2 秒間で平均して 120 個，最低 110 個のセンサーデータが取得できる．このうち，各発話データにおいて時系列で中央にある 110 個を選んでデータとしている．セクションごとにマスクを外すことによって生じる位置の変化による影響を少なくするために，収集したデータを各セクションごとに標準化して用いた．

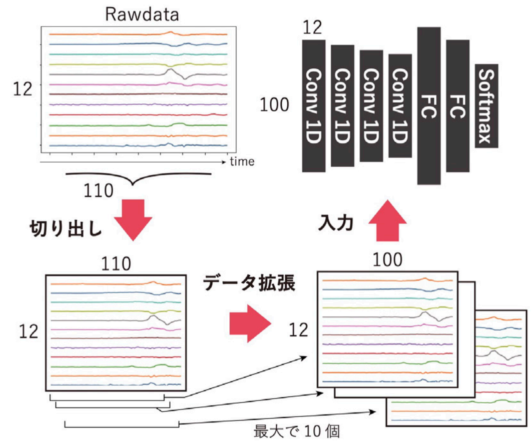

図 4　データ拡張

表 2　評価用ネットワーク 1(Conv1)

InputShape	Layer	Kernel/Stride/Pad
100 × 12	Conv1D	3/1/valid/
98 × 64	Dropout	0.5
98 × 64	Conv1D	3/1/valid/
96 × 128	Dropout	0.5
96 × 128	Conv1D	3/1/valid/
94 × 256	Dropout	0.5
94 × 256	Conv1D	3/1/valid/
92 × 512	Dropout	0.5
92 × 512	BiGRU	96
92 × 512	BiGRU	96
512	Dropout	0.5
512	FC	4096
4096	FC	22

表 3　評価用ネットワーク 2(Conv2)

InputShape	Layer	Kernel/Stride/Pad
100 × 12	Conv1D	3/1/valid/
98 × 64	MaxPool	2/none/valid
49 × 64	Dropout	0.5
49 × 64	Conv1D	3/1/valid/
47 × 64	MaxPool	2/none/none
23 × 128	Dropout	0.5
23 × 128	Conv1D	3/1/valid/
21 × 256	MaxPool	2/none/valid
10 × 256	Dropout	0.5
10 × 256	Conv1D	3/1/valid/
8 × 512	MaxPool	2/none/valid
4 × 512	Dropout	0.5
2048	FC	4096
4096	FC	22

5. 評価

5.1 評価用のネットワーク

収集したデータを評価するために 2 種類のニューラルネットワークを用意した．ネットワークへの入力は，標準化された 12 次元のセンサーデータ 100 個で，出力はコマンドを示す 22 次元の one hot vector となっている．

評価用のニューラルネットの 1 つ目 (Conv1) は Lip-Interact[5] で提案されたもの (表 2) であり．他方 (Conv2) は Derma[4] で提案されたもの (表 3) を拡張したものである．

5.2 音声コマンドの認識

収集した 2640 回 (=22 コマンド× 40 回× 4 名) のデータのうち 8 割を学習用に，残りの 2 割をテスト用に分割した後，データ拡張により 10 倍にした．学習の際の batchsize は 128 で 500epoch まで実施した．

コマンドの分類精度は表 4,5 のようになった．ユーザーごとの認識の結果の平均は，既存手法が 84.6%, 提案手法が 79.1%となった．全てのユーザーのデータでの分類は既存手法が 86.2%, 提案手法が 79.9%となった．

5.3 表情・動作の認識

収集した 960 回 (=6 動作× 40 回× 4 名) のデータを音声コマンドと同様に 8 割を学習用に 2 割をテスト用に分割した後，先述したデータ拡張により 10 倍にして用いた．

全ユーザーでの認識結果は表 6 のようになった．6 動作のみでは 84.7%，音声コマンドと合わせて 28 状態 (=22 コマンド+6 動作) での分類では 79.1%となった．

用いたネットワークは，音声コマンド・表情と動作，いずれの場合も Net2 の方が良い性能を示した．

表 4 結果 1: ユーザーごとの認識率

	net	user1	user2	user3	user4	平均
Silent Mask	Net1	0.944	0.657	0.792	0.725	0.791
	Net2	0.960	0.657	0.814	0.736	0.791
Baseline	Net1	0.834	0.705	0.909	0.873	0.831
	Net2	0.946	0.801	0.778	0.861	0.846

表 5 結果 2: 全てのユーザーでの認識率

	net1	net2
SilentMask	0.736	0.799
Baseline	0.805	0.862

表 6 結果 3: 表情・動作の分類

	net1	net2
表情・動作のみ	0.808	0.847
表情・動作 + 音声コマンド	0.756	0.790

6. 主観評価

デバイスの使用感に関して表 7 の質問を行い，リッカート尺度の 7 段階で評価した．結果を図 5 に示す．

結果として，口周辺にデバイスを直接取り付ける場合に比べて，マスク型デバイスの方が長時間装着することに対しての抵抗が少ないことが明らかになり，より日常的に用いやすいことを示した．

表 7 主観評価の質問項目

Q1	口を動かしやすかった (貼り付けた場合)
Q2	数時間 (2，3 時間) 着用できると感じた (貼り付けた場合)
Q3	1 日中着用できると感じた (貼り付けた場合)
Q4	口を動かしやすかった (マスクに付けた場合)
Q5	数時間 (2，3 時間) 着用できると感じた (マスクに付けた場合)
Q6	1 日中着用できると感じた (マスクに付けた場合)

7. 議論・今後の課題

7.1 センサーの配置

本研究では，筋電センサによるサイレントスピーチインタラクションでのセンサーの設置場所を参考に，顎と右の頬にセンサーを設置した．マスクという口周辺全体を覆えるデバイスをより活かし，センサーを複数配置した形での，口周辺や耳付近などの皮膚運動がどのように影響を与えるかは考慮できておらず，検討が必要である．また，精度の面でも複数のセンサーを配置することで口腔の無声発話によるによる皮膚運動をより精細に取得できると考えられる．

7.2 マスクのキャリブレーション

マスクは口全体を覆うデバイスであるが，皮膚とマスクの密着状態が常に維持されているわけではなく，皮膚の運動計測に変化が生じやすい．今回は環境を統一するために市販の不織布マスクを用いたが，利用するマスクの大きさは人によって異なるため，他者へのデータの適用についても考慮が必要である．また，布製のマスクなど顔の形状に合わせやすく，センサーとより密着な状態を期待できるものについても検討が必要である．

8. 結論

マスクに付着して口周辺の運動を計測する 2 つの加速度・角速度センサーにより無声発話を推定するマスク型 SSI を提案した．これによって，2 つのセンサーで取得した 12 次元の情報を深層学習を用いて解析し，22 種類の音声コマンドを 79.9%の認識率で，表情と動作の計 6 種類を 84.7%の精度で識別できた．

また，皮膚にセンサーを直接貼り付ける既存手法と比較し，同程度の性能を得た．主観評価も加えて行い，マスク型デバイスは口にセンサーを直接貼り付ける場合に比べて日常的に用いやすいことを示した．

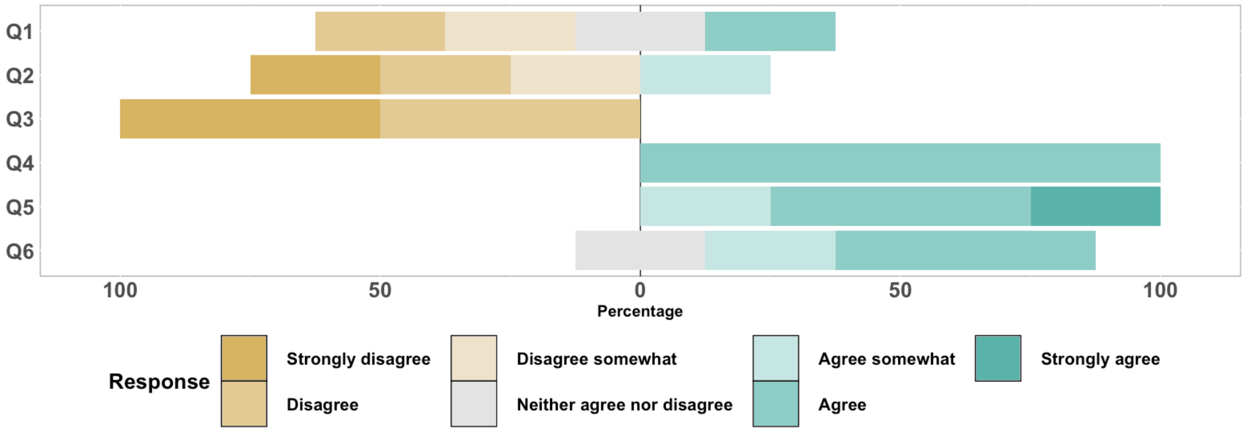

図 5　デバイスの装着に関する主観調査

マスクによる SSI は従来の手法に比べ，照明等による環境の影響が少なく，日常的に用いられるため目立たず，片手を塞ぐことがない点で優れている．また，マスクは医療用として病院等で用いられており，マスクによる SSI は患者の健康状態のデータなど，重要だが本人に秘匿すべき情報を入力する際などの応用として期待される．

参考文献

[1] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. and Brumberg, J.: Silent speech interfaces, *Speech Communication*, Vol. 52, No. 4, pp. 270 – 287 (online), DOI: https://doi.org/10.1016/j.specom.2009.08.002 (2010). Silent Speech Interfaces.

[2] Li, R., Wu, J. and Starner, T.: TongueBoard: An Oral Interface for Subtle Input, *Proceedings of the 10th Augmented Human International Conference 2019*, AH2019, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3311823.3311831 (2019).

[3] Kapur, A., Kapur, S. and Maes, P.: AlterEgo: A Personalized Wearable Silent Speech Interface, *23rd International Conference on Intelligent User Interfaces*, IUI '18, New York, NY, USA, Association for Computing Machinery, p. 43–53 (online), DOI: 10.1145/3172944.3172977 (2018).

[4] 暦本純一，西村　悠：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，インタラクション 2020，情報処理学会，pp. i–ii (2020).

[5] Sun, K., Yu, C., Shi, W., Liu, L. and Shi, Y.: Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands, *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, New York, NY, USA, Association for Computing Machinery, p. 581–593 (online), DOI: 10.1145/3242587.3242599 (2018).

[6] Wand, M., Koutník, J. and Schmidhuber, J.: Lipreading with long short-term memory, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6115–6119 (2016).

[7] Kimura, N., Kono, M. and Rekimoto, J.: SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, Association for Computing Machinery, p. 1–11 (online), DOI: 10.1145/3290605.3300376 (2019).

[8] Csapó, T., Grósz, T., Gosztolya, G., Tóth, L. and Markó, A.: DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface, (online), DOI: 10.21437/Interspeech.2017-939 (2017).

[9] Li, R., Wu, J. and Starner, T.: TongueBoard: An Oral Interface for Subtle Input, *Proceedings of the 10th Augmented Human International Conference 2019*, AH2019, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3311823.3311831 (2019).

[10] Meltzner, G. S., Heaton, J. T., Deng, Y., Luca, G. D., Roy, S. H., and Kline, J. C.: Development of sEMG sensors and algorithms for silent speech recognition., Vol. 15, No. 4, pp. 1741–2552 (2018).

[11] Wand, M. and Schultz, T.: Session-independent EMG-based Speech Recognition., pp. 295–300 (2011).

[12] Sakashita, M., Kawahara, K., Koike, A., Suzuki, K., Suzuki, I. and Ochiai, Y.: Yadori: Mask-Type User Interface for Manipulation of Puppets, *ACM SIGGRAPH 2016 Emerging Technologies*, SIGGRAPH '16, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/2929464.2929478 (2016).

[13] 綾郁石井，孝徳小松，直　橋本：HappyMouth:マスク型デバイスによる対面コミュニケーション能力の拡張，技術報告 7，明治大学, 明治大学, 明治大学 (2018).

[14] 山崎友翼，志築文太郎，　高橋伸：マスク型インタフェースによるハンズフリーな入力手法，インタラクション 2018，情報処理学会，pp. 612–614 (2018).

[15] Nam, H. Y., Hernandez, I. and Harmon, B.: Unmasked, *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20 Adjunct, New York, NY, USA, Association for Computing Machinery, p. 111–113 (online), DOI: 10.1145/3379350.3416137 (2020).

[16] Lee, H., Kim, Y. and Bianchi, A.: MAScreen: Augmenting Speech with Visual Cues of Lip Motions, Facial Expressions, and Text Using a Wearable Display, *SIGGRAPH Asia 2020 Emerging Technologies*, SA '20, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3415255.3422886 (2020).