視線入力とリップリーディングを用いたハンズフリーインタフェース

蘇 子雄[1,a)]　張 シン磊[1,b)]　木村 直紀[1,c)]　暦本 純一[1,2,d)]

概要：
視線入力は手や音声による操作を必要としないといった利点で特徴付けられており，それに基づいた様々なインターフェースが提案されている．しかし，視線計測器の性能制限や固視微動現象の影響などにより，視線入力の精度はまだ不十分である．一方で，音声入力は同様にハンズフリーな操作方式であるが，公共の場では情報漏洩や周囲に迷惑をかける他，音声認識の精度に依存するなどの制限が挙げられる．そこで本研究は，視線選択・無声発話操作を併用した手や音声によるそうさを必要としない入力方式 Gaze+Lip を提案する．従来の視線のみの入力方式との比較実験を通して，以下の三つの利点があることを検証した：第一に，視線入力の曖昧さを解消し，誤操作が生じにくい入力を可能にした；第二に，視線選択した上で無声発話で操作することによって，迅速かつ繊細な入力を実現した；第三に, 視線選択から得られたコンテキスト情報を活用し，リップリーディングの認識精度を改善した．最後に，NASA-TLX アンケートによる主観評価を実施し, ワークロードが低減されたことがわかった．

1. 背景

障害者に向けの入力装置や他の作業と同時に進行できる操作手段などとして，ハンズフリーインタフェースが多くの場面に応用されるように追求されてきた．その中で，眼球追跡技術の発展に伴い，視線入力デバイスの普及が進みつつある．人間があるオブジェクトを取り扱う前に，本能的にそのオブジェクトに視線を向ける [1]．そのため，遠隔で選択やクリックなどの操作を行う際において，視線入力は非常に迅速かつ効率的であることが示された [2]．しかしながら，視線入力インタフェースには三つの課題があると考える．まず，固視微動といった眼球運動によって人間の目の動きの測定は難しく，市販の視線計測器だと ± 1 ° の範囲では測定結果が不正確になる [3]．この不正確さは，距離が離れたスクリーン上に射影するとさらにセンチメートル級の誤差に拡大される．そのために，細小なオブジェクトを操作することは困難である．次に，視線入力は表現力に欠ける．マウスのクリック操作の代わりに，注視や瞬きなどの目の動きはよく用いられてきたが，この手法では意図しない入力が行われてしまう（”Midas Touch” issues [4]）が課題として残されている．最後に，視線入力インタフェースは緻密な制御が必要であるため，長時間の使用する場合は疲労を感じやすい [2]．

一方で，音声入力はもう一つのハンズフリーな入力方式として挙げられる．自然言語を用いてコンピューターを直感的に操作することができるため，すでに様々なスマートデバイスに組み込まれている．しかし公共の場おいて，音声によるインタラクションは情報漏洩や周囲に迷惑をかけることや，認識精度がノイズに左右されてしまう可能性がある．

そこで本研究は，視線または音声の利点を取り入れた上，視線入力とリップリーディング（Lipreading）によるマルチモーダルインタフェースを提案し，テレビ画面を遠隔で操作するプロトタイプ Gaze+Lip を実装した．提案手法では，視線入力で目標を選択した上，無声発話による細かい操作を行うため，手や音声による操作を必要としない．また，従来の視線のみによる入力方式との比較実験を設計し，定量的かつ定性的な評価を行い，提案手法の有効性を検証した．なお，本研究の成果は [5] で詳説されており，こちらも参照されたい．

2. 関連研究

関連研究はマルチモーダル視線入力インタラクションとリップリーディングを用いた無声発話インタラクションの

[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)] zxsu@g.ecc.u-tokyo.ac.jp
[b)] xinleiz@g.ecc.u-tokyo.ac.jp
[c)] kimura-naoki@g.ecc.u-tokyo.ac.jp
[d)] rekimoto@acm.org

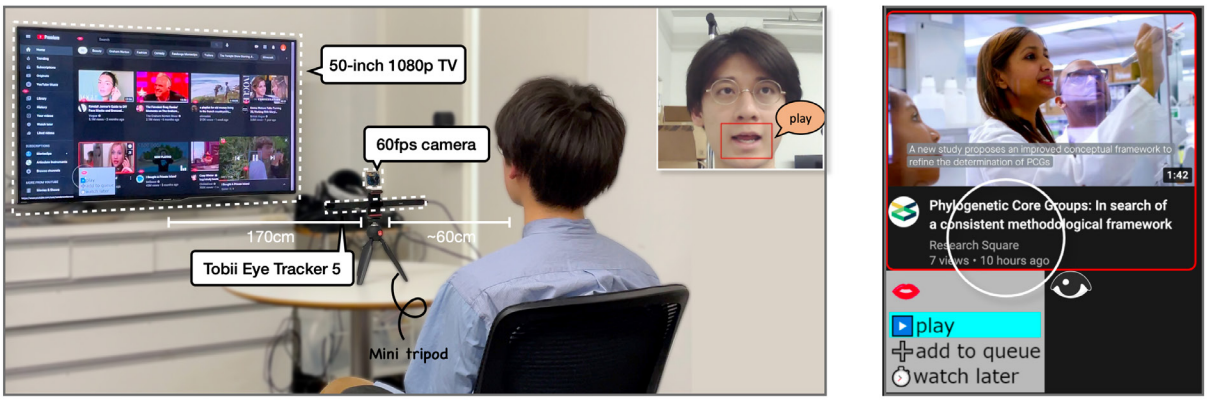

図 1　システムの構成

二つのカテゴリに分類できる．

2.1　マルチモーダル視線入力インタラクション

視線における課題をモダリティの拡張によって解決することは長年の研究テーマとなっている．その中，ジェスチャー入力は様々な環境において表現力に富んだインタラクションを提供できるため，視線入力と併用した研究が行われてきた．例えば，距離の離れたスクリーンにおいて，視線で選択したオブジェクトをジェスチャーを用いて操作する手法が考えられた [6] ．また，直接的な操作の他に，選択段階にジェスチャーを導入し，細かにオブジェクトに対して選択の精度と速度を改善した手法も提案されている [7]. [8] は，タブレットのインタフェースに着眼し，タッチジェスチャーの出力先を視線位置に変更するインタフェースを提案した．しかしながら，これらのジェスチャー入力による手法では，モーションキャプチャーの検知エリアやスクリーンの上に手を置かなければならないため，ユーザーの手の活動が束縛される他，疲労が生じやすいなどの課題が存在する．

2.2　リップリーディングに基づいた無声発話インタラクション

深層学習によるリップリーディングは，視覚情報により口唇の動きのみを用いて発話内容を認識することである．音声情報を使用しないため，発声インタラクションの欠点を回避できる．近年，深層ニューラルネットワークを用いたリップリーディングに関する研究は飛躍的に進んでおり [9], [10], [11]，様々な大規模なデータセットが作成されている [12], [13]．しかしながら，リップリーディングを用いたインタラクションに対する議論はまだ不十分である．木村ら [14] は胸部に装着するビデオカメラを用いたウェアラブルリップリーディングデバイスを提案している．Lipinteract [15] では，リップリーディングシステムをスマートフォンに実装し，タッチ操作と併用する入力方式を提案している．前述のようなリップリーディングのみを用いたインタフェースは，あくまでも一般的な入力方式を補佐するものであり，一部のシナリオにおいてシステムのショートカットとして機能するに過ぎない．さらに，唇の画像情報だけでは情報量が不十分のため，同一の口形素（viseme，唇の動きのパターンの構成単位）にからできたフレーズはほぼ不可分である．既存のリップリーディングインタフェースは精度を優先にした結果，語彙数が制限されている．（TieLent では 15 語，Lipinteract では最大 10 語）．

上記の課題を解決するために，本研究では，視線入力とリップリーディングによる無声発話を相補的な手段として結合し，Gaze+Lip と呼ばれるマルチモーダルな入力インタフェースを提案する．

3.　Gaze+Lip：視線入力とリップリーディングを用いたマルチモーダルインタフェース

3.1　デバイス構成

図 1 にシステムの全体構成を示す．ユーザーは 50inch（112cm×65cm）のテレビから 230cm 離れた位置に着座した状態で操作する．Tobii 社の Tobii Eye Tracker 5 [16] により取得したテレビ画面上の注目点を視線入力として用いられる．リップリーディングは市販のビデオカメラで得られた映像（フレームレート 30，サイズ横 800 x縦 600 ピクセル）により実装する．視線計測器とビデオカメラはミニ三脚で固定し，常に被験者の正面 60cm の位置から撮影するように設置される．取得されたデータは，直接接続している Windows 10 PC により処理し，ユーザーの入力結果をテレビの画面に反映する．

3.2　視覚情報による発話検出

In the wild の実験を通じて実環境における性能を評価するため，視覚情報による発話検出を行い，発話時のみの口唇映像を抽出する．具体的に，ビデオカメラにより正面から撮影された映像の各フレームは，Dlib [17] を用いて顔のランドマーク推定を行う．推定されたランドマークに基づいて，口唇領域の座標と開口角度を計算する．特に，開口

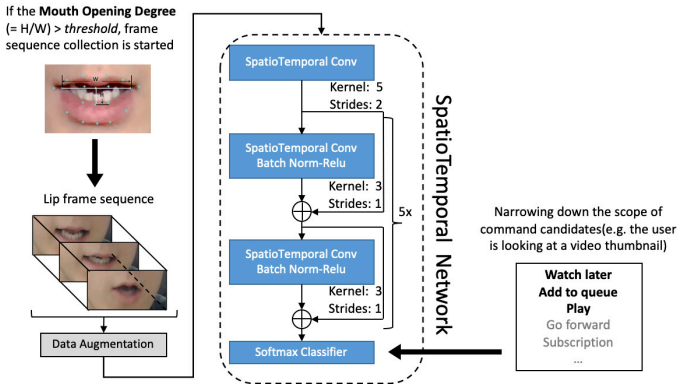

図 2 リップリーディングに用いるニューラルネットワークの構成

表 1 Gaze+Lip に用いた 27 種類のコマンド

watch later	add to queue	homepage	go back	go forward
scroll up	scroll down	trending	subscription	library
original	play	stop	previous	next
expand	volume up	volume down	caption	full screen
like	dislike	share	save	delete
notification	profile			

角度は上唇と下唇の隙間の高さと幅の比率として定義される．さらに，実験の準備段階では，被験者ごとに違う閾値（算出方法は後節により説明される）が決定され，実験中ではこの閾値とリアルタイムの開口角度を用いてユーザーが発話しているかを判定する．被験者の開口角度が連続 20 フレーム閾値を超えると口唇領域にクロップされた映像の録画が開始し，開口角度が連続 20 フレーム閾値 T を下回ると録画が終了する．録画開始する前の 20 フレームも保存される．

3.3 エンドツーエンドリップリーディング

リップリーディング（視覚情報によるスピーチ認識）のために，スペイシオウテムパラル・コンボリューション・ネットワーク（Spatiotemporal Convolution Network, SCN）”R（2+1）D” [18] に基づいた方法を用いた．”R（2+1）D”では，3 次元畳み込みを 2 次元畳み込みと 1 次元畳み込みに分解することによって計算量を抑えつつ，空間と時間の特徴を有効に抽出することができる [19]．そのため，リップリーデイングを含めた動作認識のタスクによく使われている [15], [20]．本研究で使用したネットワークの構成は図 2 に示される． 3.5 で収集されたフレームシーケンスは時間（t）× 高さ（40）× 幅（80）× チャンネル（3）の Tensor として入力され，Spatiotemporal Conv 層にダウンサンプリングされた後に 5 つの Spatiotemporal Residual Block によって 1024 次元のベクトルにエンコードされる．最後に，このベクトルは softmax 関数により分類される．

3.4 サイレントスピーチコマンドの選択

スマートテレビの遠隔操作という日常的な応用シナリオを想定し，YouTube のウェブサイトを閲覧するときによく使われる 27 個のコマンドを認識対象として設定する（表 1）．

3.5 データ収集

構築した認識手法を評価するために，6 人の実験参加者（女性 5 人，男性 1 人）を募集した．すべての参加者は視線や無声発話による入力インタフェースを使用した経験を有してない．リップリーディングモデルの学習データを収集するため，各参加者に 27 個のコマンドを無声で発話させ，データ記録する．一回につき一つのコマンドがテレビ画面上に提示され，参加者はそれを見ながら無声で読み上げるように指示される．コマンドの提示される順番はランダムであり，画面上における出現位置も毎回変わる．これによって顔の傾きの影響を取り入れ，モデルの汎化能力を向上させる．データ収集の準備段階では発話検出に用いる開口の閾値の測定を行う．参加者は発話しない状態でテレビ画面の四隅をそれぞれ 5 秒間注視し，その間の開口角度の最大値 $\mathcal{M}$ が記録される．開口角度の閾値 $\mathcal{T}$ は値 $\mathcal{T} = \mathcal{M} + 0.02$ として決められる．参加者は，発話を正確にできた場合は Space キーで保存して次のコマンドに進み，間違えた場合はキーボードの R キーでやり直すことができる．27 個のコマンドを 1 セットとし，各セット終了後に 1 分間の休憩を挟み，計 18 セットを行った．最終的に，各参加者から 486（$= 27 \times 18$）回の発話データを収集した．

3.6 リップリーディングモデルの学習

収集されたデータを用いて，各参加者ごとのリップリーディングモデルを訓練した．さらに，各フレームシーセンスに以下の 5 種類のデータオーグメンテーション手法を適用いた：Random Time Shift, Random Rotation, Random Resize, Random Color Jitter, Random Frame Drop．一回分の学習は batchsize 8 で 20 epoch まで実施し，Nvidia RTX 2080Ti GPU ボード 2 枚で分散学習を行う場合，所要時間約 10 分である．実験で使われる Nvidia GTX 1080 搭載の Windows 10 PC で推定した結果，Dlib による顔認識とリップリーディング両方の処理時間は合わせておおよそ 134ms である．提案した認識手法を評価するために，Cross Validation 法（fold=9，すなわち訓練データと検証データの比率は 8:1）を用いて各参加者のデータセット上で精度検証を行った．98.42% ± 1.11% の分類精度が達成できた．

4. 比較実験

本節では，従来の視線のみを用いた Gaze-Dwell 方式との比較実験を行い，以下の三つのリサーチクエスチョンを軸として調査する：Gaze-Dwell に比べて，Gaze+Lip は表現力の高いサイレントスピーチコマンドを導入したが，RQ1. 入力速度に影響は与えるか．RQ2. より繊細な操作

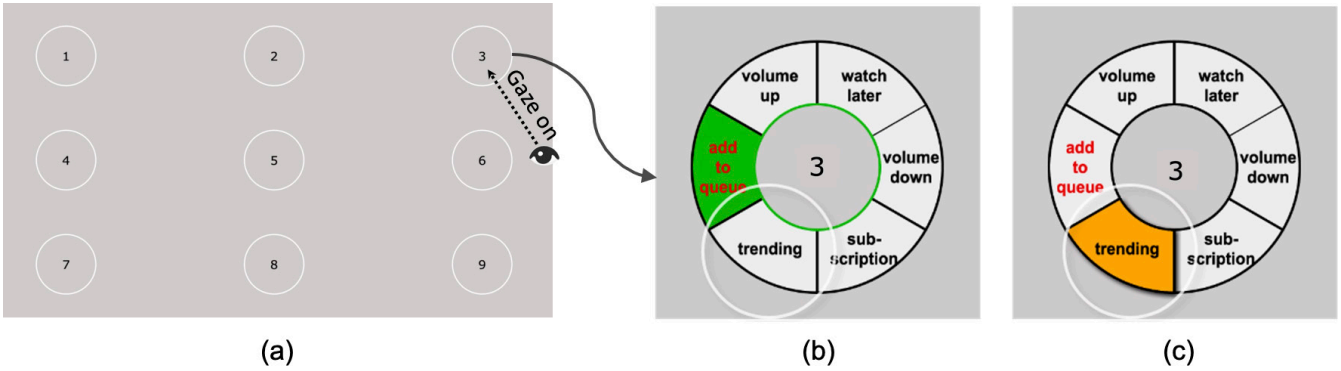

図 3 比較実験に用いたパイメニュー選択タスクの GUI

は実現できるか．RQ3. 誤操作を抑えられるか．これらの疑問を答えるために，タスクの完成時間，タイムアウト率，誤操作の回数の三つの評価指標を測定し，その結果に基づいた解析を行う．

4.1 実験条件

図 3 で示す比較実験用の GUI を開発した．画面上に 9 つのパイメニューが均等に配置されており，視線位置を表すカーソル（白い円形）を重ねると 6 個のボタンが周りに展開されるが，カーソルが外れるとまた閉じてしまう．27 個のコマンドから無作為に選出された 6 個のコマンドがそれぞれボタンの上に提示され，その中の一つだけが赤文字になっている．このコマンドに対応するボタンを目標とし，Gaze-Dwell 方式，あるいは Gaze+Lip 方式でクリックすることを被験者に行わせる．メニューが展開された瞬間に 5 秒間のカウントダウンが開始し，時間内にタスクを完成できなかった場合は「タイムアウト」として記録され，直ちに次の試行に移行する．その間に間違ったボタンをクリックしたことを誤操作として記録される．Gaze-Dwell と Gaze+Lip それぞれの仕組みは以下のように説明する：

Gaze-Dwell 方式：ボタン領域に視線が 500 ms 以上に停留することでクリックする．この方式では，間違ったボタンに視線を停留することによって，誤操作が生じうる．さらに，メニューが閉じないようにするため，操作中には視線カーソルをメニューから離してはならない．

Gaze+Lip 方式：ボタンに表示されているコマンドを無声発話で読み上げることによってクリックする．この方式では，メニューを視線で展開すると可能なコマンドを確定できるため，リップリーディングの認識候補をそのメニューに含まれるコマンドたけに絞ることができる．その結果，認識精度の上昇することが期待できる．参加者が口を開けると同時に，展開されたパイメニューがロックされ，無声発話の間には視線を自由に動かすことができるため，注視時間が短縮される．

4.2 実験手順

評価における順序効果を相殺するために，Gaze+Lip と Gaze-Dwell のどちらを先に行うかについ 2 通りの提示順序を用意し，参加者を 2 つのグループに分けてそれぞれ別の順序で実験を行った．各入力条件における実験は以下の手順で行った．最初に，各参加者にデモンストレーションを用いて 2 種類の入力方式を説明し，3 分間の自由練習を実施する．次に，アイトラッカーのキャリブレーションを行い，パイメニューによるタスクを行う．全部の 27 個のコマンドが一回ずつ目標コマンドとして設定され，これらの試行を完成すると 2 分間の休憩をとる．休憩中には，NASA-TLX [21] のアンケートに回答してもらう．

5. 実験結果

5.1 定量評価

各参加者から入力方式 2 条件 ×27 試行=54 回の実験データが得られた．このデータから算出した各参加者平均のタスクの完成時間，タイムアウト率，誤操作の回数を評価指標とした分差分析を行った．（表 2）．タスクの完成時間（$F_{1,6} = 4.45$, p = .061）について，平均値で比較すれば Gaze-Dwell 方式のほうが入力時間が短いが，有意差が認められなかったため，Gaze+Lip は Gaze-Dwell と同等の速度で入力する可能性が示唆された（RQ1，RQ2）．一方で，タイムアウト率（$F_{1,6} = 11.84$, p = .0063），誤操作の回数（$F_{1,6} = 5.61$, p = .039）について Gaze+Lip の平均値が低く，有意差が認められた．したがって，Gaze+Lip はより繊細な操作を可能にし，誤操作も抑えられたことがわかった（RQ3）．さらに，どの指標においても提案手法は標準偏差が低かったことによって，ユーザーに依存しない性能有しているが示唆された．

5.2 主観評価

NASA-TLX の結果を図 4 に示す．主観的作業負荷が高いほど測定値が高くなる．平均値で比較する場合，全ての項目において Gaze+Lip の作業負荷が低いことが観測され

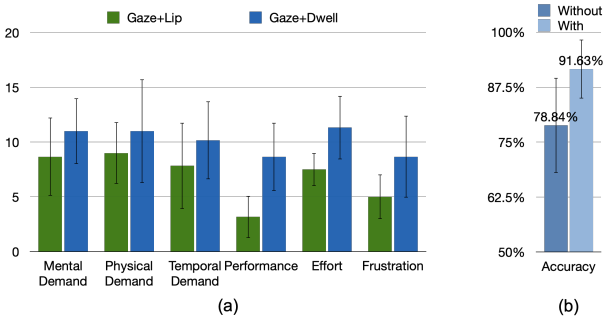

図 4 （a）NASA-TLX によるワークロードの主観評価の結果．（b）視線入力を用いてリップリーディングの認識候補を絞り込むことによる精度改善の結果．

表 2 比較実験で有効性を示すために用いる評価指標 ($M \pm SD$)．

入力方式	タスク完成時間（ms）	タイムアウト率	誤操作の回数
Gaze+Lip	2290.27 ± 121.59	3.09% ± 4.330%	1.67 ± 1.506
Gaze-Dwell	1876.70 ± 464.76	25.93% ± 15.538%	7.67 ± 6.022

た．特に，Effort（$F_{1,6} = 5.32$, p = .044）と Performance（$F_{1,6} = 8.71$, p = .015）の項目では有意差が認められた．この結果は，Gaze+Lip ではより少ない努力を要求されながら，より高い作業成績を達成できることを示唆している．

6. 議論・今後の課題

マルチモーダリティによるリップリーディング精度の改善

前節で行った比較実験の他に，視線入力によりリップリーディングの認識範囲を限定したことによる精度の改善について議論する．そのため，実験中のリップリーディング出力を記録し，リップリーディングのみを使った場合，すなわち候補コマンドを 27 種類そのまま絞り込まない場合の認識率も算出した．その結果を図 4（b）に示す．視線入力のコンテキスト情報を活用したことよって，リップリーディングの認識率の平均値が 12.79%上昇したことがわかった．視覚情報には曖昧性が固有しており，高精度のリップリーディングは困難である．そのため，語彙に制限のないサイレントスピーチインタフェースは未だに実現されていない．本研究得られたマルチモーダリティに関する知見は，無声発話の表現力を拡張する新たな可能性を提示した．

無声発話の検出

本研究では，開口角度による無声発話の検出を行い，その角度がある閾値を超えたときにリップリーディングが動作する仕組みになっている．しかし，参加者が不意に口を開けてしまう場合や普通に会話するときは誤操作を起きてしまうといった課題が残されている．解決策としては，音声認識に用いられるキーワードスポッティングを模倣した特定なリップジェスチャーによって始めるリップリーディングシステムなどが興味深い将来課題である．

7. 結論

視線入力によって目標を選択し，さらに無声発話で細かい操作を行うハンズフリーインタフェースを提案した．実証実験を通して，提案手法の方が従来の視線のみによる入力方式より正確かつ繊細なインタラクションが可能であることを示した．さらに主観評価の結果は，作業負担が全体的に低いことを示唆した．特に努力 (Effort) と作業達成度 (Performance) の項目において，本手法の性能が顕著に優越であることが認められた．視線入力と無声発話を併用した提案手法は，手や音声による操作を必要としない前提で迅速，正確かつ簡単に使用できるため，ハンズフリーインタフェースの応用場面をより一層広げることが期待される．

参考文献

[1] Sibert, L. E. and Jacob, R. J. K.: Evaluation of Eye Gaze Interaction, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, New York, NY, USA, Association for Computing Machinery, p. 281–288 (online), DOI: 10.1145/332040.332445 (2000).

[2] Zhai, S., Morimoto, C. and Ihde, S.: Manual and Gaze Input Cascaded (MAGIC) Pointing, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, New York, NY, USA, Association for Computing Machinery, p. 246–253 (online), DOI: 10.1145/302979.303053 (1999).

[3] Nevalainen, S. and Sajaniemi, J.: Comparison of three eye tracking devices in psychology of programming research., *PPIG*, Vol. 4, pp. 151–158 (2004).

[4] Jacob, R. J. K.: What You Look at is What You Get: Eye Movement-Based Interaction Techniques, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, New York, NY, USA, Association for Computing Machinery, p. 11–18 (online), DOI: 10.1145/97243.97246 (1990).

[5] Su, Z., Zhang, X., Kimura, N. and Rekimoto, J.: Gaze+Lip: Rapid, Precise and Expressive Interactions Combining Gaze Input and Silent Speech Commands for Hands-free Smart TV Control, *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–6 (2021).

[6] Zhang, Y., Stellmach, S., Sellen, A. and Blake, A.: The costs and benefits of combining gaze and hand gestures for remote interaction, *IFIP Conference on Human-Computer Interaction*, Springer, pp. 570–577 (2015).

[7] Chatterjee, I., Xiao, R. and Harrison, C.: Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions, *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, New York, NY, USA, Association for Computing Machinery, p. 131–138 (online), DOI: 10.1145/2818346.2820752 (2015).

[8] Pfeuffer, K., Alexander, J., Chong, M. K. and Gellersen, H.: Gaze-Touch: Combining Gaze with Multi-Touch for Interaction on the Same Surface, *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, New York, NY, USA, Association for Computing Machinery, p. 509–518 (online), DOI: 10.1145/2642918.2647397 (2014).

[9] Assael, Y. M., Shillingford, B., Whiteson, S. and

de Freitas, N.: LipNet: Sentence-level Lipreading, *CoRR*, Vol. abs/1611.01599 (online), available from ⟨http://arxiv.org/abs/1611.01599⟩ (2016).

[10] Wand, M., Koutník, J. and Schmidhuber, J.: Lipreading with Long Short-Term Memory, *CoRR*, Vol. abs/1601.08188 (online), available from ⟨http://arxiv.org/abs/1601.08188⟩ (2016).

[11] Chung, J. S., Senior, A. W., Vinyals, O. and Zisserman, A.: Lip Reading Sentences in the Wild, *CoRR*, Vol. abs/1611.05358 (online), available from ⟨http://arxiv.org/abs/1611.05358⟩ (2016).

[12] Chung, J. S. and Zisserman, A.: Lip Reading in the Wild, *Asian Conference on Computer Vision* (2016).

[13] Chung, J. S., Senior, A., Vinyals, O. and Zisserman, A.: Lip Reading Sentences in the Wild, *IEEE Conference on Computer Vision and Pattern Recognition* (2017).

[14] Kimura, N., Hayashi, K. and Rekimoto, J.: TieLent: A Casual Neck-Mounted Mouth Capturing Device for Silent Speech Interaction, *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3399715.3399852 (2020).

[15] 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).

[16] AB, T. P.: Tobii Pro Lab.

[17] King, D. E.: Dlib-ml: A Machine Learning Toolkit, *Journal of Machine Learning Research*, Vol. 10, pp. 1755–1758 (2009).

[18] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M.: A Closer Look at Spatiotemporal Convolutions for Action Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[19] Ji, S., Xu, W., Yang, M. and Yu, K.: 3D Convolutional Neural Networks for Human Action Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp. 221–231 (online), DOI: 10.1109/TPAMI.2012.59 (2013).

[20] Prajwal, K., Mukhopadhyay, R., Namboodiri, V. P. and Jawahar, C.: Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13796–13805 (2020).

[21] Hart, S. G. and Staveland, L. E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, *Advances in psychology*, Vol. 52, Elsevier, pp. 139–183 (1988).