KuchiPaKu:サイレントスピーチインタラクションを学ぶための ゲーム的教材の研究

酒井ちひろ[†1]　　大島登志一[†1]

サイレントスピーチインタラクション（SSI）の基本原理と効果について学べるゲーム的学習教材の開発を行う．現代社会では会話において静粛性・秘匿性を維持しなければならない場面があり SSI が注目されている．そのような状況下で SSI を学びたいというニーズがあるが，それに応えるための教材は開発されていない．これを解決するために誰でも簡単に SSI を学ぶことができる教材を開発したいと考えた．ユーザは口を動かすだけで PC 画面上の対象を操作し，同時に処理過程を見ることで簡単に SSI の概要と原理を学ぶことが可能である．

1. はじめに

本研究では「サイレントスピーチインタラクション（SSI:Silent Speech Interaction）」の基本原理と効果について学べるゲーム的学習教材の研究を行う．ユーザは口を動かすだけで PC 画面上の対象を操作し，同時に処理過程を見ることで SSI の原理と効果をわかりやすく学ぶことが可能である．医療現場では発声できない患者とコミュニケーションが取れないことが課題となっている．この課題を解決するために SSI を活用した発声支援デバイスが開発されている．このような SSI の社会的な実装を普及させるためには SSI を専門に研究している人以外でも原理と効果を理解する必要がある．しかし，現時点でそれらを記載した教材は開発されていない．そこで本研究ではこの課題を解決するために技術の応用を考える技術者と医療関係者を対象とし，わかりやすく SSI の原理と効果の理解を促すゲーム的な教材，KuchiPaku を開発することとした．本論文では試作中のシステムについて説明する．

2　関連研究

2.1SSI の医療への応用

医療現場では，発声できない患者とコミュニケーションがとれないことが課題となっている．これを解決するために様々な発声支援デバイスが開発されている．喉頭摘出後の患者のコミュニケーション手段として 1959 年ごろから使用されているのが，電気式人工喉頭（EL: Electrolarynx）である．円筒状のデバイスで，本体ボタンを押すことで先端から振動が発生し，咽頭原音の代わりとなる振動音を電気的に与えている．しかし使用する際に片手が塞がる，機械的な音声しか発生することができないなど問題点が存在する．

これらの課題を解決するための発声支援デバイスに応用されているのが SSI である．例えばマサチューセッツ工科大学の Kapur らはユーザの顔に電極を貼り付け，神経筋信号を読み取る Alter Ego というデバイスを開発した[1]．このデバイスは神経筋信号を入力データとしているため，口を開いて言葉を発しない場合でも淀みない会話が可能である．また森川らは喉頭摘出者のための歌唱支援システムの構築を目指し，電気式人工喉頭を用いて得られる電気音声をより自然な歌声に変換する手法を提案した[2].

本研究では発生せずともコミュニケーションが可能な SSI に着目し，前述した高度な研究を支える基礎知識を学習することができるツールを試作することにした．また SSI の効果である口を動かすだけでインタラクションが可能であるという点をゲーム的なインタラクションを通じてユーザに理解してもらうために，対戦型ゲームをコンテンツに組み込む．

2.2 音声変換機構の関連研究

暦本は話者・非言語依存の自己教師型学習に基づく実時間ささやき声音声変換機構，WESPER を開発した[3]．この音声変換機構を使用することにより囁き声から変換された音声の品質が向上し，韻律の自然さも保持されることが確認された．小泉らは固定点反復に基づくノイズ除去拡散確率モデル（WaveFit）のような反復フレームワークに GAN（GAN: Generative Adversarial Network）のエッセンスを統合した，WaveFit というニューラルボコーダを開発した[4]．このフレームワークにより推論速度が向上し，さらに主観的なリスニングテストでは WaveFit によって合成された音声と人間の自然な音声の間に統計的に有意な差はみられなかった．このような音声変換機構の利用は本研究の次の段階で検討しており，検出した口の形からユーザ自身の通常音声を再現することが最終目標である．

2.3 口腔形状センシングと囁き声の収音の関連研究

Li らは TongueBoard というリテーナ型デバイスを発し，静電容量式タッチセンサを口蓋に配置することで舌の動きをセンシングした[5]．囁き声の収音方法としては福本が開発した SilentVoice が挙げられる[6]．この研究では呼気では

†1 立命館大学
Ritsumeikan University

なく吸気を音源とするイングレッシブ・スピーチ方式を採用しており，ポップノイズに悩まされることなく収音することが可能である．このように様々な音声変換手法や口腔形状認識手法が研究されており，口腔形状をデバイスにより直接検出する方法は本研究の次段階において検討中である．

3　提案手法

3.1 システムの基本デザイン

本研究では様々な SSI の手法の中でも共通する基本原理である母音の認識に焦点を当てる．子音の認識に関してはマイクロフォンを装着したり，顔にデバイスを貼付したりする必要があり母音の認識に比べて応用的な技術が必要となるが，基本原理の学習をユーザに促すことを目的としているため本研究では子音の認識については取り扱わない．ユーザが理解する必要がある SSI の基本原理と効果，それに対する提案手法は以下の通りである(表 1)．またシステム全体の流れを以下に示す(図 1)．USB カメラでリアルタイムの映像を取得し，顔と口の検出を行う．検出した口の高さと幅のアスペクト比からユーザが意図している母音を推定する．推定された母音に対応するコマンドに応じて画面内の操作対象が動くという流れになっている．また口がカメラで検出されていること，口の幅と高さが母音推定に活用されていること，母音に入力コマンドが割り当てられていることをそれぞれユーザに理解してもらうためにデバッグモニタを 3 つ表示する（図 2)．

	ユーザの学習内容	提案手法
SSIの基本原理	口腔形状認識の仕組み	口腔形状を正面からカメラで撮影し、処理内容のデバッグ画面をPC画面上に表示
	母音推定の仕組みとその活用	口の高さと幅のアスペクト比から母音を推定するデバッグ画面を表示し、PC画面上の操作対象が母音ごとの入力コマンドに応じて動く
SSIの効果	医療への応用 新しいコンピュータインタラクションツールとしての活用	ユーザへ文字情報として提示

表 1　ユーザの学習内容と提案手法

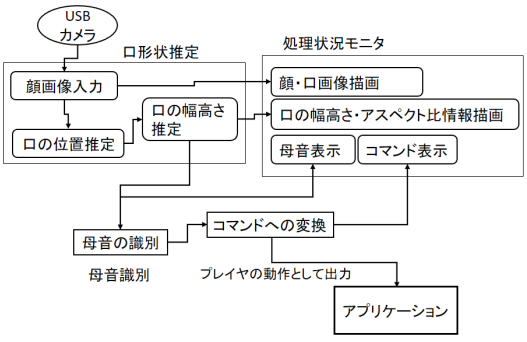

図 1　実装中の機能ブロック図

3.2 コンテンツデザイン

本研究で取り扱うゲーム的コンテンツの実装について説明する．第一にユーザは SSI の基本原理として口形状認識の仕組みと母音推定の仕組み・その活用を学習する．ユーザは声を出さないことを前提条件としているため，顔に装着するマイクロフォンを利用せず，コンピュータビジョンのみで口形状認識を行う．認識は Web カメラで行う．認識した口形状のリアルタイム映像から口の位置推定を行い，口角と上・下唇の中央点の 4 点を特徴点として検出し，ユーザが意図している母音の推定を行う．ユーザが意図している母音は特徴点から得られた口の高さと幅のアスペクト比から推定する．この際ユーザに口形状認識の仕組みと母音推定の仕組みを理解してもらうために，顔・口の検出モニタ，口の幅と高さ，そのアスペクト比の数値を表示するモニタ，母音と対応するコマンドの表示モニタをスイッチで切り替えて PC 画面上に表示できるようにする（図 2)．

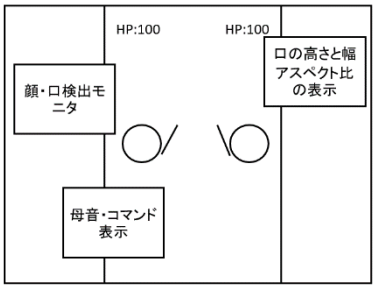

図 2　ユーザ体験画面のイメージ

そして推定された母音にあらかじめ設定されている入力コマンドに応じ，PC 画面上の操作対象が動く仕組みである．第二にユーザは SSI が医療へ応用できることや，新しいコンピュータインタラクションツールとして活用できるということを SSI の効果として学習する．これらはユーザが体験手順を見ながら学習できるよう文字情報として提示する．

4　システムデザイン

4.1 システム構成

- 言語　C++
- 画像処理・認識ライブラリ　OpenCV Ver 4.1.1

なお実装中の機能ブロック図（図 1)に倣い，それぞれのブロックのシステムを説明する．

4.2 口形状認識

廣瀬らの研究を参考とし，ユーザの顔と口の検出は OpenCV のライブラリのオブジェクト検出を用いる[7]．今回は OpenCV の顔と口のカスケードファイル「haarcascade_frontalface_default」と「haarcascade_mcs_mouth」を用いる．これらを用いることによりリアルタイムでユーザ二人の顔と口の位置検出を行う．なおカスケード分類器のみで口の検出を行うと誤検出が多いため，検出された顔の下 4 分の 1 の範囲からのみ口の検出を行うという検出範

囲の制限を設けた．次に制限範囲内から口の検出を行い，口の特徴点抽出を行う．廣瀬らの研究を参考に，エッジ抽出を行ったのちに OpenCV ライブラリの goodFeaturesToTrack を用いて特徴点の検出を行う．エッジ抽出フィルタはリアルタイム性の維持と精度を考慮し Sobel フィルタの使用を試みる．また特徴点抽出に関して廣瀬らは口角の 2 点を抽出していたが，本研究では口の高さと幅を検出するために下唇と上唇の中心も特徴点とし，合計 4 点検出する．

4.3 処理状況モニタ

ユーザに母音推定の仕組みについて理解を促すために処理状況デバッグモニタを画面上に表示する．表示するモニタは 3 つある．第一に顔・口検出モニタである．これは検出されているユーザの顔と口を表示するモニタである．リアルタイムで取得しているカメラの映像に検出されている顔と口を囲むように矩形を描画したものを表示する．これによりユーザは顔と口が検出されていることを理解し，同時に正しくそれらが検出されているかどうかを矩形によって確認することができる．第二に口の高さと幅の表示モニタである．ユーザの口の幅と高さ，幅と高さのアスペクト比を数値として画面上に表示することにより母音の口の形と数値の対応付けを体感してもらうことができる．第三に母音・コマンド表示画面である．ここでは推定された母音とそれに対応するコマンドを文字情報として表示する．これによりユーザが意図している母音とコンピュータが推定する母音が一致しているかどうかをユーザ自身が確認することができる．これら 3 つの画面はスイッチでユーザが任意で表示できるようにする．

4.4 母音の識別

検出した口の幅と高さのアスペクト比からユーザが意図している母音を推定する．推定する母音とそれに対応する入力コマンドは以下のようになっている．（図 3）

母音	入力コマンド
あ	進む
い	戻る
う	上に行く
え	下に行く
お	攻撃

図 3 母音と対応する入力コマンド

4.5 コマンドへの変換

推定された母音とそれに対応する入力コマンドによりユーザは操作対象に特定のコマンドを入力することができる．ユーザ二人がそれぞれ自身の操作対象にコマンドを入力し，体力メータ（HP）が 0 になったユーザは負けという設定になっている．

5 考えうる課題点

本研究で考えうる課題点はシステムの遅延である．口の検出と操作対象の操作を同じスレッドで行うとリアルタイム性を維持できない可能性がある．その解決策としてマルチスレッド化が挙げられる．顔と口の検出・特徴点抽出・母音推定をサブスレッド，操作対象の操作をメインスレッドで行うことで遅延を防ぐ手法を検討する．

6 まとめ

本研究は SSI の基本原理と効果をユーザにわかりやすく伝えるためのゲーム的教材の開発を目指している．ターゲットである医療従事者や技術者が SSI について学ぶことにより，SSI が医療へ応用される機会が増える可能性がある．本研究では医療従事者と技術者を対象とした SSI の基本原理と効果の理解を促すゲーム的な教材の開発を目指すため，コンピュータビジョンを用いた口形状認識と対戦型ゲームをコンテンツとしたインタラクションの実現を行う．しかしコンピュータとのインタラクションしか想定していないことや母音しか認識することができないことが現状の課題である．今後の方針としてはコンピュータだけでなく人とのインタラクションを組み込むことや子音も認識できるようにすることが実用的なコミュニケーションツールを実現させる第一段階であると考える．

現在医療現場では発話が困難な患者と医師のコミュニケーションが課題になっている．SSI が発声支援デバイスもしくは新しいコミュニケーションツールとして応用されることによりこの課題を解決できる可能性がある．より多くのユーザが SSI について学び応用を考えることで医療現場での課題が少しでも解決されることを望んでいる．今後は自身が SSI を応用した新たなコミュニケーションツールや発生支援デバイスを開発し，患者と医師の円滑なコミュニケーションを支援したいと考えている．

謝辞

本研究は，JSPS 科研費 JP21K12004 の助成を受けたものです．

参考文献

1) Kapur, A., Kapur, S., & Maes, P. (2018, March). Alterego: a personalized wearable silent speech interface. Proc. of 23rd International conference on intelligent user interfaces (pp. 43-53)..

2) 森川一穂, 戸田智基: “喉頭摘出者のための歌唱支援を目指した統計的電気音声変換法”, 研究報告音楽情報科学 (MUS)2017.27 (2017): pp.1-6.

3) Rekimoto, J. (2023, April). WESPER: zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions. Proc. of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-12).

4) Koizumi, Y., Yatabe, K., Zen, H., & Bacchiani, M. (2023, January).

WaveFit: an iterative and non-autoregressive neural vocoder based on fixed-point iteration. Proc. of In 2022 IEEE Spoken Language Technology Workshop (SLT) (pp. 884-891). IEEE.
5) Li, R., Wu, J., & Starner, T. (2019, March). Tongueboard: an oral interface for subtle input. Proc. of the 10th Augmented Human International Conference 2019 (pp. 1-9).
6) Fukumoto, M. (2018, October). Silentvoice: Unnoticeable voice input by ingressive speech. Proc.of the 31st Annual ACM Symposium on User Interface Software and Technology (pp. 237-246).
7) 廣瀬明依, 孫 氷玉, 宮中 大, 早川 吉彦: “タブレット端末による非接触咀嚼検出アプリの開発”, 医用画像情報学会雑誌 33.3 (2016): pp.57-62.