K-024

日本語発話訓練システムのための画像及び音声処理の試行
Trial of image and voice processing for Japanese speech training system

松下 開星† 荒平 高章†
Kaisei MATSUSHITA Takaaki ARAHIRA

キーワード：日本語発話，画像処理，発話訓練システム
Keywords：Japanese speech, Image processing, Speech training system

1.はじめに

先天性難聴や元は聴力があったが，何らかの不慮の事故などにより聴力を失った中途失聴者といった聴覚障がいを有する人々にとって，言語の発声・発話訓練は極めて重要なことである．しかしながら，音声によるフィードバックを得ることが難しいため，自身の発話が正確かどうかを把握することや他者と比較することが難しく，訓練が困難であるという課題がある．特に日本語の発音は，口の動きの違いが微細であり，視覚的にその差異を判別することが困難である．そのため，聴覚以外の情報に頼らざるを得ない場面では，認識が難しくなる傾向がある．

Lipnet ネットワークを活用したもの[1]，発話シーンからのキーフレーム検出に基づく手法[2]，発話映像の各フレームに対する口形の類似度を特徴量パラメータとして利用したもの[3]，顔画像情報と音声情報を統合したもの[4]といった情報処理技術を用いて読唇を実現させる研究が行われており，加えて，人工知能（AI）技術の発展に伴い，それら音声認識や画像処理を応用した発話支援システムの研究も進められている．中でも，先ほどの Lipnet や MediaPipe などのライブラリを用いた顔のランドマーク抽出は，カメラ映像から口元の動きを高精度に捉えることが可能であり，視覚的な発話フィードバックの提供に有効であると考えられる．

発話訓練システムの例としてコトリハ（KOTOREHA）という発語失行などの言語障害を持つ方のために開発された言語支援アプリケーションがある．これは口型動画・音声・絵カードを用いたトレーニングにより，視覚と聴覚の両方から発話を促進することができる．また，家庭での自主練習が可能で，個別の単語登録や達成度の可視化にも対応しており，言語聴覚士不足や訓練機会の制限を補う支援ツールとして有効であると考えられる．コトリハの場合，事前に提示された単語に従って模倣する形式で訓練が行われるが，発話のお題が与えられない自由発話の状況では，発音の正誤や改善点を利用者自身で把握することが難しく，訓練効果が限定的となる可能性がある．

本研究では，顔画像から得られる口周辺のランドマーク情報を用いて，日本語の母音を識別する AI モデルの構築を試みる．最終的には，視覚的に自身の発話を確認・学習できる支援システムの基盤を構築することを目的とする．

2.方法

今回は，訓練システム開発の前段階として日本語母音での発話訓練システムの試行を行った．

2.1 環境

本研究では，以下の開発環境及びライブラリを使用した．

・言語：Python
・実行環境：Windows 11（Visual Studio Code）
・使用ライブラリ：
—Open CV
—MediaPipe
—PyTorch

2.2 MediaPipe

MediaPipe とは，Google が開発したクロスプラットフォーム対応の機械学習推論およびメディア処理フレームワークであり，顔，手，身体などの高精度なランドマーク検出をリアルタイムで実行可能とする．中でも Face Mesh モデルは，顔全体から 468 点（拡張時は 478 点）の 3 次元ランドマークを抽出できる．今回は，Face Mesh モデルを用いて口元や顎を含む顔下部のランドマークを抽出し，発話に伴う口の動きの定量化に活用した．以下に示す画像は実際に MediaPipe を利用し，ランド―マークを描画している画像である．

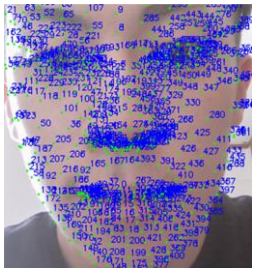
図 1 描画の様子

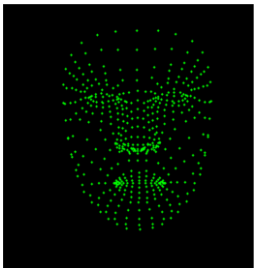
図 2 描画される点

2.3 実験対象者

「あ」「い」「う」「え」「お」と閉口形である「ん」で認識を行った．4 名（17～23 歳）に協力してもらい母音と閉口形をそれぞれ 50 シーンずつ撮影した．本研究では以下のように A～D で被検者を割り当てる．

・男性 A（日本人）
・男性 B（日本人）
・女性 C（日本人）
・女性 D（日本人）

† 九州情報大学院経営情報学研究科経営情報学専攻

2.4 ランドマーク取得

MediaPipe により抽出されたランドマークは，正規化された座標（0〜1）で表現されており，画像サイズに応じてスケーリングすることでピクセル単位の座標に変換される．今回，発話の瞬間に撮影した各ランドマークのインデックス，x 座標，y 座標を取得し，対応する静止画像ごと保存した．

保存されたデータには，撮影時の 1 フレームのランドマーク座標が記録されており，これを機械学習の特徴量として利用するため，以下のような前処理を行った．

まず，特定の発話状態（例：「あ」や「ん」など）ごとにデータを分類し，それぞれの母音ごとと閉口形の六つに分けて整理した。各データから x 座標，y 座標の数値のみを抽出し，1 サンプルあたり [x1, y1, x2, y2, ..., xN, yN] の形に並べて 1 次元ベクトルに変換する．加えて，各サンプルに対応するラベル（母音クラス）を付与することで，教師あり学習に適したデータセットを構築した．

学習には PyTorch を用い，入力次元を「使用ランドマーク数 × 2」とした全結合ニューラルネットワークを構成し，母音分類を目的とした多クラス分類タスクとして訓練を行った．

2.5 実行

学習済みのニューラルネットワークモデルを用い，実際に Web カメラ映像からリアルタイムで取得した顔画像に基づいて，母音の発話認識を行った．

MediaPipe によって得られる顔のランドマークから，口周辺および顎の領域に属する点を抽出し，それぞれの x，y 座標を取得する．取得したランドマーク座標は，学習時と同様に[x1, y1, x2, y2, ..., xN, yN] の形式で 1 次元ベクトル化され、モデルへの入力とされた．

モデルは前処理済みのベクトルを受け取り，各母音クラス（例：「あ」「い」「う」「え」「お」「ん」）に属する確率を出力する．最も確率が高いクラスが，現在の発話状態として分類される．

出力された結果は画面上に母音ラベルとして表示され，訓練者は自身の口の形や発音状態に応じたフィードバックを視覚的に確認することができた．

3.結果と考察

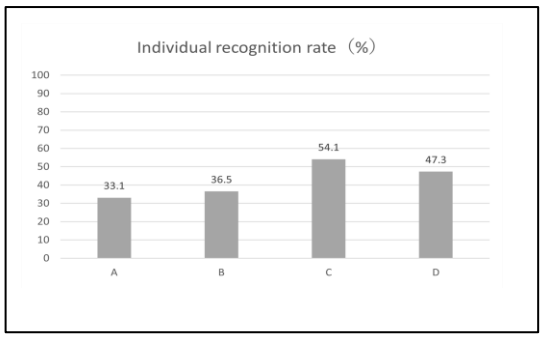

図 3　個人の認識率

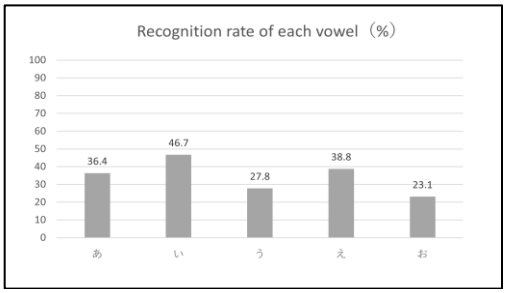

図 4　各母音の認識率

各個人の認識率は図 3 のようになった．性別で見ると A, B が男性であり，C，D が女性であるが，これを比較すると女性のほうが男性よりも認識率が高かった．各口形の認識率は図 4 のような結果となった．全体的に認識率が低く，特に「う」と「お」に関してはどちらも三割を下回ってしまった．

今回の結果から発話者の口形の明瞭さと認識率の高さは比例関係にあり，実際にそれが顕著に出る結果となった．特に「う」と「お」は発話時の口形が非常に似ていたため，どちらもいずれかに認識されてしまうという結果が多かった．結果を見て全体的に識別精度に限界があることが分かり，主な要因として母音ごとの口の動きが類似している場合にランドマークの変化量が小さく、誤分類が発生しやすいことと，カメラの角度や被写体の姿勢変化の影響で、同一母音であっても撮影条件によりデータのばらつきが生じた可能性があることが考えられる．

4.今後

本研究では，日本語発話訓練システムの開発を目指し，システムを運用していくにあたっての前段階である母音認識を行った．画像処理を日本語の母音で行い，口の開き方による精度のブレ，「う」，「お」の母音の類似などの問題点がみられた．母音認識では満足のいく結果を得ることができなかったため，次には日本語単語の認識のためのデータを収集し，また精度向上とシステム開発，音声処理機能の実装に努めていく．そして，訓練システムの単語認識が完了次第すぐに第三者，対象者の方々にもテスト，評価をしていただき，フィードバックを受け取って本格的に形にしていきたいと考えている．提案する訓練システムの対象者は，聴覚障がいや発話障がいといった言葉を発言することが困難，発言できていたとしても比較的不明瞭な部分が見受けられるといった方々としているため提案するシステムは，ユーザビリティを重視し，手軽に使いやすい訓練システムを目標に開発していきたい．

参考文献

[1] 北原　瑠伊，張　力峰，”機械学習を用いた日本語読唇における五十音データセット作成の提案”，産業応用工学会全国大会 2021 講演論文集

[2] 齊藤　剛史，森下　和敏，小西　亮介，”発話シーンからのキーフレーム検出とキーフレームに基づく単語読唇”，IEEJ Trans. EIS，Vol.131，No.2(2011)

[3] 宮崎　剛，中島　豊四郎，”口形ベースの機会読唇における単語認識手法の提案と評価”，DICOMO2014

[4] 奥村　晃弘，濱口　佳孝，岡野　健治，宮崎　敏彦，”顔画像情報と音声情報の統合による発話認識”，情報処理学会論文誌，Vol.39 No.12(1998)