

## D-44 画像による読唇技術のための機械学習を用いた母音認識の検討

岸和田樹 福本尚生 伊藤秀昭

(佐賀大学理工学部)

## 1. はじめに

読唇技術は、騒音下での音声認識補完や聴覚障がい者のコミュニケーション支援、無音映像の解析など多様な応用が期待されている。日本には約36万人の聴覚障がい者がおり、音声だけでは情報を得にくい場面も多く、日常生活で不便を感じている人も少なくない。唇の動きなど視覚的手がかりを活用した日本語読唇技術は、今後の支援手段として大きな可能性を持つと考えられる。近年、英語の読唇技術はディープラーニングの発展によって、著しい進歩を遂げている[1]。一方、日本語は声調言語であり、口唇の動きだけでは識別が困難な音も多く、依然として難易度が高い。

本研究では、MATLABとWebカメラを用いてリアルタイムに口元を検出し、ディープラーニングで母音を認識する日本語読唇システムの構築を目指す。具体的には、YOLOX-TinyモデルをMATLAB上で実装し、口領域を検出した。次に、映像特徴から母音を認識し、音声を用いずに発話内容を推定した。

## 2. 口領域検出モデルの構築

リアルタイムで読唇を行うためには、まず安定した口領域の検出が不可欠である。そこで、約260枚の顔画像をWebカメラで撮影し、データを収集した。収集した画像には、MATLABのImage Labelerを用いて各画像の口領域に「Mouth」ラベルを付与した。このラベル付きデータを用いて、YOLOX-Tinyモデルによる機械学習を行い、口検出モデルを構築した。学習済みモデルをWebカメラからの画像に適用することで、リアルタイムに口領域を検出するシステムを実現した。

その結果、正面を向いた顔に対しては高精度かつ安定的に検出できるモデルを構築することができた。一方で、横顔や被写体が遠くにある場合には、検出精度の低下が見られた。

## 3. 母音認識の検証

母音の認識を行うために、まず/A/, /I/, /U/, /E/, /O/に加えて口を閉じている状態/X/の計6種類の口の画像約240枚を収集した。データはYouTubeの動画やWebカメラを用いて取得し、各クラスに手動で分類し整理した。その後、データ数を補うために左右反転、回転、コントラスト調整などの拡張処理を行い、最終的に6クラスあわせて約1800枚の画像を用意した。これらの画像にラベルを付与し、機械学習による分類モデルを構築した。

学習済みのモデルはWebカメラからのリアルタイム映像に適用し、各フレームから推定される母音クラスを表示することで、リアルタイムな母音認識を実現した。実際の認識例を図1に示す。黄色の四角は口領域を示し、その領域で認識した母音とその信頼度を表している。

次に、リアルタイムで/A/, /I/, /U/, /E/, /O/, /X/に対して、対応する口の形を提示し、モデルが最初に認識したラベルと比べ、正答判定を行った。各発話について10回ずつ試行し、正解率を算出して認識精度を評価した。その結果を表1に示す。また、各クラスについての10回分の信頼度の平均を表2に表す。



図1 各母音のリアルタイム認識

表1 各ラベルの正解率

各ラベル	/A/	/I/	/U/	/E/	/O/	/X/
正解率(%)	100.0	100.0	70.0	100.0	90.0	20.0

表2 各母音信頼度の平均

各ラベル	/A/	/I/	/U/	/E/	/O/	/X/
平均(%)	92.3	94.1	87.7	94.1	90.3	58.8

その結果、/A/, /I/, /E/など口の開き方に特徴のある母音は高い精度で認識することができた。一方で、/U/, /O/など類似した口の形を持つ母音では、認識精度がやや低下する傾向が見られた。さらに、/X/は/U/, /O/と誤って認識されるケースが多く、6クラスの中でも特に精度が低い傾向を示した。しかし、リアルタイム映像の各フレームから母音を推定する簡易手法であっても、一定の認識性能を得られることが確認でき、今後の改善次第では実用的なリアルタイム読唇システムへの応用が期待される。

## 5. まとめ

本研究では、日本語の母音を視覚情報だけでリアルタイムに認識するシステムの構築を目的として、まずYOLOX-Tinyを用いた口検出モデルを開発した。このモデルはWebカメラ映像に適用することで、実時間で安定的に口領域を抽出することが確認できた。その後、/A/, /I/, /U/, /E/, /O/, /X/の6クラスに対する母音認識を行い、約1800枚の静止画データを用いた分類モデルの学習と、リアルタイムで推論結果を出力するシステムの実装に成功した。

今後は、より高精度な母音認識を実現するため、現在の静止画ベースの認識から一歩進めて、動画データを用いた時系列モデルの導入を検討していきたい。口の動きの時間的変化を学習に取り入れることで、動的な特徴を活かした認識精度の向上が期待される。データの追加収集や拡張処理の強化、さらにネットワーク構造の工夫も併せて検討していく。これらの取り組みによって、実用的で高精度なリアルタイム読唇システムの実現を目指す。

## 5. 参考文献

- [1] YEWELI XIAO et al, "Lip Reading in Cantonese" in IEEE Access, vol. 10, pp. 95020-95029, 2022