日本語読唇におけるマルチモーダル手法

Multimodal approach for Japanese lip reading

西川岳都
Gakuto Nishikawa

1　はじめに

現在、日本の障碍者手帳を持つ総人数は 416 万人を超えており, そのうち聴覚・言語障害者の数は 38 万人とされている [1]. 高齢者を中心に耳が聞こえにくくなり, 聞き取りが困難になったり, 聞き間違いが多発するようになっている. それだけではなく また, 音声による会話では発音がはっきりしていなかったり, 声の特徴によっては聞き取りづらい場合がある. そのため, 口唇の動きから発話内容を推定する”読唇”技術が発展すれば, より円滑な会話が可能になると考えた.

読唇技術が向上することで, 聴覚に障害のある人々とのコミュニケーションを支援できるだけでなく, 失声症などの理由で声を発することができない人々や聴覚障碍を持ちながらも手話を習得していない人々の会話の補助としても有用である. しかし, 日本語は母音や子音の種類が少なく, 口唇の形のみから発話内容を推定することが困難であるため, 読唇技術の研究は発展途上にある.

また, 近年様々な問題に対して情報を統合させて学習・推論を行う, マルチモーダルな手法について検討が行われている. 読唇についてもこのマルチモーダルな手法を用いることで読唇技術の発展へと寄与できるのではないかと考えた. そこで、本研究ではマルチモーダルな手法に着目した日本語読唇について検討する.

2　関連研究

読唇に関する研究はこれまでも行われている. イギリスのオックスフォード大学や GoogleDeepMind の研究者らによる共同チームが開発した LipNet は, 動画フレームとテキスト情報を利用したモデルである [2]. 読唇術の専門家が事前に決められた文章を発話する英語話者の読唇を行った場合の正解率は 52.3 %であった. それに対し, LipNet の正解率は 95.2 %であった.

日本語の読唇についても同様に行われている. 駒井らは Active Appearance Model で抽出した口唇の特徴点から音素の認識を行った [3]. その認識精度は母音で 67.81 %, 子音で 11.85 %となった. また, 柿原らの CNN を用いたマルチモーダル音声認識の手法では, 画像特徴量のみを用いた場合, 認識精度は 50.9 %になった [4]. しかし, こちらの実験では使用される単語が 2620 単語と少なく, 評価用に用いられた単語もわずか 216 単語である. また, どちらの手法でも認識精度は低く, 完全な読唇はできていない.

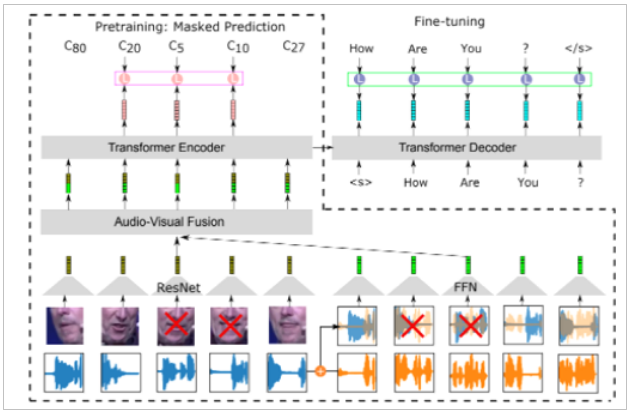


図 1: AV-HuBERT のアーキテクチャ [5]

3　提案手法

本研究では，Meta 社が 2022 年に提案した音声・映像の自己教師あり学習モデルである AV-HuBERT[5] を用い，事前学習済みモデルに Transformer ベースのデコーダを接続し，読唇認識タスクに対してファインチューニングを行った.

AV-HuBERT は，音声と映像の両方を活用するマルチモーダルな自己教師あり学習モデルであり，BERT に着想を得たマスク予測タスクにより事前学習される. 具体的には，音声を 39 次元の MFCC 特徴量に変換した後，k-means 法でクラスタリングを行い，擬似ラベルを生成する. その後，音声は 26 次元の対数フィルタバンクエネルギーを FFN に, 映像データは 3D-ResNet に通して特徴量を抽出し，それらの一部をマスクした入力を Transformer エンコーダに与え，マスクされた部分のクラスタを予測させる. また, これらの中間表現をもとにクラスタを作り直し, 再度作り直したクラスタで学習することでモデルの精度を高めていく.

本研究では，AV-HuBERT の事前学習済みエンコーダ（Transformer 24 層，隠れ次元 1024，マルチヘッドアテンション数 16）を用い，デコーダとして Transformer 12 層，隠れ次元 1024，アテンションヘッド数 8 のモデルを新たに構築した. ファインチューニングでは，エンコーダの重みを初期化に用いつつ，映像のみを入力とし

一般社団法人情報処理学会, IPSJ
関西学院大学大学院, Kwansei Gakuin University Graduate School

て単語系列を出力するように学習を行った.

学習には，自作の映像・テキストペアから成る日本語データセットを用い，最終的な性能評価として Word Error Rate（WER）を指標として計測した.

データセットは，評価用データ 100 件とテスト用データ 100 件をランダムに抽出し，残りを学習用データとする 10 個の学習用データセットを構成した.

学習実験としては，以下の 2 通りの条件でファインチューニングを行った．実験 1 は，前処理した映像とテキストをそのまま AV-HuBERT に入力し，ファインチューニングを行う通常の方法である．実験 2 は, 映像の単語数を 2 → 4 → 6 → 8 →全文と段階的に増やした方法である．各段階において 10 エポックずつ学習を行い, モデルを段階的に訓練した.

4　使用したデータセット

データセットは 2 種類の日本語映像・字幕データセットを作成した.

1 つ目のデータセットは，京都大学の料理インタビュー対話コーパス（CIDC）および国立情報学研究所の IDR データセット提供サービスにより大阪大学から提供を受けた「大阪大学 マルチモーダル対話コーパス Hazumi」に収録されたインタビュー映像を利用したものである [6][7]. 使用した映像データの総量は約 100 時間で，発話単語数は約 73 万語、ユニーク単語数は 7735 単語である. 以下これをデータセット 1 とする.

2 つ目のデータセットは，ニュース映像を自ら撮影・収集し，手動で字幕を作成したものである．会話データと比較してフィラーや曖昧な発話が少ない. こちらのデータセットも使用した映像データの総量は約 100 時間で, 発話単語数は約 101 万語, ユニーク単語数は 12500 語である．以下これをデータセット 2 とする.

これらのデータセットは，dlib を用いて話者の口唇領域を検出し，88 × 88 ピクセルのグレースケール映像として切り出した．また，字幕テキストには表記ゆれの修正などの前処理を加え，映像と正確に対応づけたデータセットを構築した.

テキストデータ
いや全然大丈夫です私も最後らへんには思い出しとこう
うーんありがとうございます
お友達とランチはいお姉ちゃんと一緒にランチに行ったりします

図 2: 作成したデータセットの口唇映像（上）とテキストデータ（下）[5]

5　実験結果

表 1, 1 に実験 1, 実験 2 におけるそれぞれのデータセットのテストデータでの WER を示す．実験 1 におけるデータセット 1 の WER の平均は 103.1 %, データセット 2 の WER の平均は 102.7 %, 実験 2 におけるデータセット 1 の WER の平均は 96.6 %, データセット 2 の WER の平均は 94.2 %となった.

	データセット 1 の WER	データセット 2 の WER
	88.8 %	90.5 %
	103.7 %	105.3 %
	104.1 %	103.8 %
	106.2 %	101.6 %
	111.0 %	108.6 %
	112.3 %	89.2 %
	90.3 %	113.5 %
	105.7 %	104.7 %
	103.2 %	102.4 %
	105.7 %	106.9 %
平均	103.1 %	102.7 %

表 1: 実験 1 におけるデータセット別の WER と平均の WER

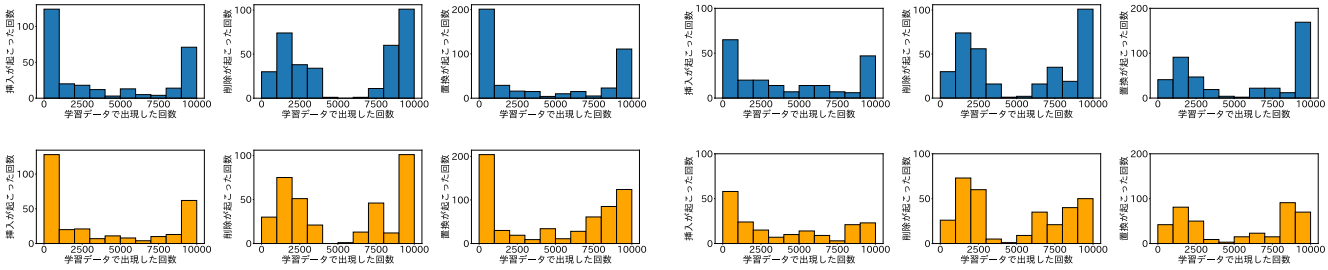
	データセット 1 の WER	データセット 2 の WER
	83.2 %	84.5
	94.6 %	95.3 %
	101.3 %	101.5 %
	100.5 %	94.2 %
	99.4 %	93.2 %
	98.6 %	89.2 %
	85.6 %	98.3 %
	101.3 %	93.5 %
	104.5 %	97.7 %
	97.4 %	94.8 %
平均	96.6 %	94.2 %

表 2: 実験 2 におけるデータセット別の WER と平均の WER

6　考察

図 3 は、実験 1 および実験 2 におけるそれぞれのデータセットの学習データ中の単語の平均出現回数と, テストデータ中における挿入・削除・置換の平均回数を示したグラフである．この結果から, すべてのモデルにおいて WER が非常に高く, 日本語の読唇精度が極めて低いことがうかがえる．したがって, 本実験においては, データの質による認識精度の差はあまり顕著ではないことがわかる.

実験 1　　実験 2



(a) 挿入　(b) 削除　(c) 置換　(a) 挿入　(b) 削除　(c) 置換

図 3: 実験 1(左)・2(右) におけるデータセット 1（上）・2（下）の学習データにおける挿入・削除・置換の平均頻度

また, 図 3 に示すように, 学習データ中で出現回数が少ない単語は, 挿入・削除・置換のいずれの観点においても誤り回数が多くなっている. 一方で, 出現頻度が高い単語についても誤り回数が多い傾向が見られる. これは, 出現頻度の高い単語ほど学習が進んでおり, 逆に他の単語の誤りとして置き換えや挿入対象になっていることが示唆される. 一方で, 出現頻度が高いにもかかわらず削除が多く発生している単語がある点は, 単語の学習がうまく進んでいない可能性を示している.

実験 1 と実験 2 を比較すると, どちらの結果においても WER は依然として高いものの, 両データセットともに実験 2 ではわずかに WER が改善していることが確認できる. これは, 実験 2 では単語単位での処理が可能となり, 単語の境界をある程度捉えられるようになったためと考えられる. 一方, 実験 1 では文全体を入力としているため, どこからどこまでが特定の単語に対応するかの判別が難しく, 認識性能が低下していると考えられる. また, 両データセットの結果において, 実験 2 では実験 1 と比較して挿入された単語の数が少なくなっていることも確認できる. 単語の区切りがより正確に捉えられるようになり, 同一単語が繰り返し挿入されるといった誤りが減少したためであると考えられる.

実験 1・2 のいずれにおいても, それぞれの実験で用いたデータセット 1 およびデータセット 2 を比較すると, WER および挿入・削除・置換された単語の分布に大きな差は見られなかった. このことから, データセット間の違いによって読唇精度が大きく変化するわけではないことがわかる. 今回使用した各データセットはそれぞれ約 100 時間分の音声映像データで構成されているが, 元論文でファインチューニングに用いられていたデータ量は約 433 時間であり, 本実験ではそのおよそ 1/4 のデータしか使用できていない. そのため, 今後はこれらのデータを統合したデータセットを構築するか, さらなるデータ収集が必要である. さらに, 日本語の言語的特性も精度に影響を及ぼしていると考えられる. 英語は 13 個の子音と 26 個の母音で構成され, 口唇の動きが大きいのに対し, 日本語は 16 個の子音と 5 個の母音で構成されており, 口唇の動きが比較的小さい. これにより, 異なる単語であっても同じような唇の動きを示すことが多く, それが高い WER の一因になっていると推察される.

7　まとめ

本研究では, マルチモーダル手法に日本語データを与えることでの読唇手法について検討を行った. 2 つのデータセットについて AV-HuBERT を日本語でファインチューニングし, そのままのテキストデータを使った実験と、単語で区切った実験の 2 つを実行した. その結果, 実験 1 におけるデータセット 1 の WER の平均は 103.1 %, データセット 2 の WER の平均は 102.7 %, 実験 2 におけるデータセット 1 の WER の平均は 96.6 %, データセット 2 の WER の平均は 94.2 %となり, どちらの実験においても日本語の読唇精度が非常に低いことが分かった. しかし, 実験 1 と実験 2 の比較からデータを単語ごとに区切って徐々に増やすことで読唇精度の改善が見られた.

今後は口唇映像をさらに収集し, ファインチューニングを行うことや事前学習モデルの作成, また, 認識精度を上げるためにエンコーダやデコーダの変更などに取り組んでいきたいと考えている.

参考文献

[1] 内閣府. 令和 6 年版障害者白書.

[2] Assael, Y. M., B. Shillingford, S. Whiteson, and N. De Freitas (2016). Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*.

[3] 駒井祐人, 宮本千琴, 滝口哲也, and 有木康雄 (2010). 唇領域の aam を用いた発話認識における画像特徴量の音素解析. **第** *13* **回画像の認識・理解シンポジウム** *(MIRU2010)*, 1771–1778.

[4] 柿原康博, 滝口哲也, 有木康雄, 三谷信之, 大森清博, and 中園薫 (2015). Convolutional neural network を用いた重度難聴者のマルチモーダル音声認識. **日本音響学会講演論文集**, 197–200.

[5] Shi, B., W.-N. Hsu, K. Lakhotia, and A. Mohamed (2022). Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

[6] 岡久太郎, and others. (2022). ウェブ会議システムを利用した料理インタビュー対話コーパス. **言語処理学会 第** *28* **回年次大会**.

[7] 駒谷和範, and 岡田将吾, 大阪大学 マルチモーダル対話コーパス hazumi, 国立情報学研究所情報学研究データリポジトリ. (データセット) (2022).
URL `https://doi.org/10.32130/rdata.4.1`