長岡技術科学大学大学院
工学研究科修士論文

題目

CNN を用いた口元画像の
視点方向変換による
斜め視点リップリーディングに関する研究

指導教員
杉田 泰則　　准教授

著者
16314889　　小梶 金志郎

提出期日
令和 2 年 2 月 7 日

ABSTRACT

Oblique view-point lip reading with frontal conversion using CNN

Author : Kinshiro Kokaji
Supervisor : Yasunori Sugita

Lip reading is to understand speech using only visual information of the shape and movement of a person's lips. Automation of lip reading enables automatic generation of subtitles for video, information transmission support in environments where audio transmission is difficult, and the like. In addition, speech recognition performance can be improved by using it together with speech information

Recently, the performance of machine-learning-based lip reading has been greatly improved by the availability of large datasets and the application of neural network-based models using deep learning. However, those studies have only considered frontal or near-frontal faces. There is a problem that recognition accuracy is greatly reduced for non-frontal inputs. As a method of improving non-frontal recognition accuracy, there is a method of using images from multiple directions as a learning data set. However, there are few published data sets with multiple directions.

This paper proposes a method of using front conversion by CNN as preprocessing in order to adapt the models trained by only frontal faces to the input of non-frontal faces. In the proposed method, the required data are dataset for lip reading only from the front viewpoint and dataset for front conversion. Since the front conversion data need only to be photographed from multiple directions, data set collection is facilitated.

In the proposed method, images of the face taken from angles of 0, 30, 45, and 60 ° were used as a data set. The network used is GAN having two networks, a Generator for generating a front image and a Discriminator for identification. Competition between the two networks improves generation accuracy. The generator's loss function is a combination of pixel loss and symmetry loss in addition to adversarial loss of GAN.

Learning of the network was performed in two patterns. One was to learn frontal transformation from each angle by a different network. The other was to learn frontal transformation from multiple angles by one network.

In the experiment, a lip reading model that classifies five Japanese vowels using only frontal data was learned for evaluation of the proposed method. For verification, data with angles of 0, 15, 30, 37.5, 45, 52.5, and 60 ° were used. The recognition accuracy was

compared for the case with and without the proposed method. The recognition accuracy was significantly reduced for non-frontal inputs when the proposed method was not used. On the other hand, by applying the proposed method, recognition accuracy was improved for non-frontal inputs. Therefore, it was confirmed that frontal transformation using CNN is usefulness as preprocessing for the lip reading model considering only the frontal viewpoint.

目次

第 1 章

はじめに

1.1 研究背景

リップリーディングは、人の唇の形や動きの視覚情報のみから話の内容を推定する技術である。コンピュータを使ったリップリーディングの自動化の研究が 1980 年代から行われている [1]。リップリーディングの自動化は様々な応用先が考えられ、映像の字幕自動生成、音声伝達が困難な騒音下での情報伝達支援などが挙げられる。また、音声認識と併用することで認識パフォーマンスを向上させることもできる。

一般的には行われてきたリップリーディングでは、画像処理を用いて唇の領域、外周形状、内周形状等の検出と特徴量抽出を行い、その時系列データを元にマッチングをすることで認識を行う [2] [3]。

近年では、大規模なデータセットの利用と深層学習を適用した研究が盛んである。DeepMind(Google の AI 開発部門) とオックスフォード大学との研究では、BCC テレビ放送から数千時間にも及ぶ動画を用いて、12 万程度の英語文章のデータセットで学習を行い、文字認識率で 60.5%、単語認識率で約 50% を記録した [4] [5]。これはプロのリップリーダの文字認識率が約 41.3%、単語認識が約 26% という記録を上回っている。また、このモデルでは従来の単語レベルでの検出でなく、文章レベルでの精度も向上している。日本語のリップリーディングについてもいくつかの研究が行われており、音素の認識において約 48% の認識率の報告がある [6]。

これらの深層学習モデルでは、CNN(畳み込みニューラルネットワーク) や LSTM のような回帰型ニューラルネットワークの適用により飛躍的な精度の向上が成された。しかし、多くの研究で正面視点での認識のみを対象としており、斜め視点などの非正面の入力に対して認識精度が低下する問題がある。Chung らの研究 [7] では、正面のみでなく複数視点を含んだ約 15,000 単語からなるデータセットを用いて学習することで、単語認識率、文字認識率ともに斜め視点で約 5%、真横視点で約 20% の精度向上が報告されてい

る。しかし、公開されているリップリーディングの大規模なデータセットは正面視点か正面付近 (30° 程度) がほとんどである [7] [8]。また、大規模なデータセットになると英語のテレビ映像がメインである。そのため、この手法ではデータセット収集の難しさとデータ数の増加等の問題がある。

1.2　研究目的

本論文では、正面視点のみを考慮して学習されたリップリーディングのモデルを非正面視点の入力に対応させることを目的として、CNN による正面変換を前処理に用いる方法を検討する。必要となるデータセットが正面視点のリップリーディング用データと正面変換用のデータになり、リップリーディングの学習で非正面のデータが必要なくなる。作成が困難であるリップリーディング用のデータセットに比べて、正面変換用データセットは、発音時の画像を複数視点から撮影するだけで良いため、データセット収集が容易になる。よって、提案法を用いたリップリーディングの実験を行い、提案手法の有用性を評価する。

1.3　本論文の構成

本論文の構成は以下の通りである。第 1 章では、本論文の研究背景と目的を述べた。第 2 章では、提案手法で用いるニューラルネットワークの基礎理論について述べる。第 3 章では、提案手法である CNN を用いた正面変換について、データセットとネットワークモデルの説明を行う。第 4 章では、提案手法を評価するためのリップリーディングのネットワークについての説明とそれを用いた提案手法の有用性の評価を行う。第 5 章では、本論文についてのまとめと今後の課題について述べる。

第 2 章

基礎理論

2.1　ニューラルネットワーク

ニューラルネットワークは、人間の脳内にある神経細胞 (ニューロン) がシナプスの結合により形成する神経回路網を数学的モデルで表現したものである。学習により、シナプスの結合強度を変化させ問題解決能力を持たせることができる。

ニューラルネットワークの例を図 2.1 に示す。ここで、左の列から入力層、中間層、出力層と呼ぶ。図中の丸はニューロン (またはノード) を表し、入力信号が閾値を超えると値を出力する。また、矢印はシナプスを表し、この結合強度により情報の伝達しやすさが変わる。

入力信号の総和に対するニューロンの出力を決定する閾値を決める関数を活性化関数と呼ぶ。活性化関数を $h()$ とすると、図 2.1 における y_1 の出力の計算は式 (2.1) で表され、各入力と結合強度の積の総和に活性化関数を適用したものになる。

$$y_1 = h(w_{11}x_1 + w_{21}x_2 + w_{31}x_1) \tag{2.1}$$

活性化関数としてよく用いられるものとしては ReLU(Rectified Linear Unit) 関数がある。ReLU 関数は式 (2.2) で表されるように入力が 0 を超えていればそのまま出力し、0 以下なら 0 を出力する関数である。

$$h(x) = \begin{cases} 0 & (x \leqq 0) \\ x & (x > 0) \end{cases} \tag{2.2}$$

ReLU は、計算の単純さによる計算の速さと $x > 0$ において、微分値が常に 1 のため、誤差逆伝搬の際の勾配消失の心配がないメリットがある。

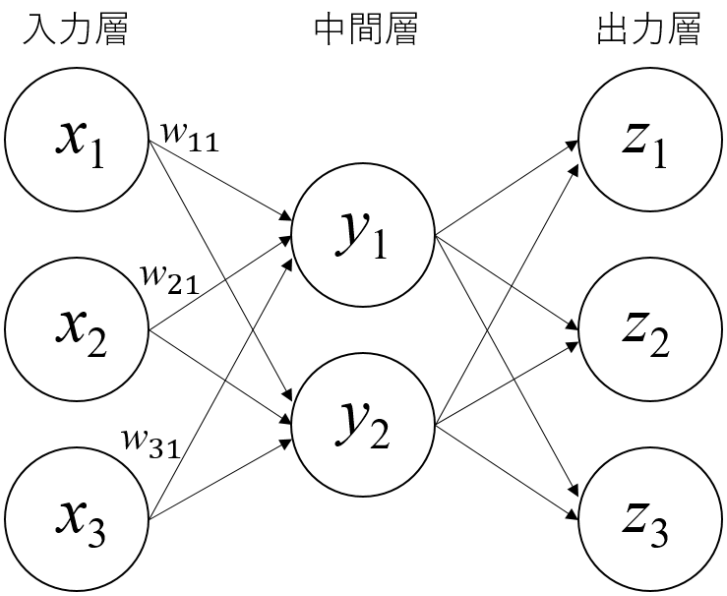

図 2.1　ニューラルネットワークの例

2.2　CNN（畳み込みニューラルネットワーク）

畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) とは、Convolution 層 (畳み込み層) をもつニューラルネットワークのことである。2010 年から始まった ILSVRC(ImageNet Large Scale Visual Recognition Challenge) という画像認識のコンテストにおいて、2012 年に CNN を用いた AlexNet [9] というモデルが優勝して以降、画像の特徴抽出がディープラーニングで行えることが示され、画像の分野において CNN が大きな注目を集めた。

通常のニューラルネットワークでは、隣接する層のすべてのニューロン間で結合が行われている。そのため、画像等を入力とした際の縦・横・チャンネル方向の 3 次元データを 1 次元データにする必要があり、データの形状が無視される。一方で、CNN は入力データを 3 次元データとして受け取り、3 次元データとして出力する。そのため、画像などの空間的特徴をもつデータを正しく利用できる可能性がある。

2.2.1　Convolution 層

Convolution 層 (畳み込み層) の基本構造を図 2.2 に示す。入力をチャンネル数 C、高さ H、幅 W のデータ (C,H,W)、フィルタをチャンネル数 C、高さ FH、幅 FW の (C,FH,FW) の場合を考える。このとき、入力とフィルタのチャンネル数は等しくなければならない。入力に対してフィルタによる畳み込み演算処理を行う。畳み込み演算は、図 2.3 に示すよ

うに、対応する位置のフィルタの要素を入力の要素を乗算し、和を求める処理を行う。これをフィルタの位置をスライドさせながら計算する。この畳み込み演算により、入力のどこに特徴が存在するかを示したものが生成され、これを特徴マップという。この特徴マップを活性化関数に通したものが出力となる。

図 2.2 では、フィルタ 1 つに対して (1 ,OH,OW) の特徴マップが生成された。N 種類のフィルタを用いることで、(N,OH,OW) の特徴マップが生成される。また、畳み込み演算時のフィルタのスライドする間隔をストライドといい図 2.2 の例では 1 だが、ストライドが 1 より大きいと入力より特徴マップのサイズが小さくなる。

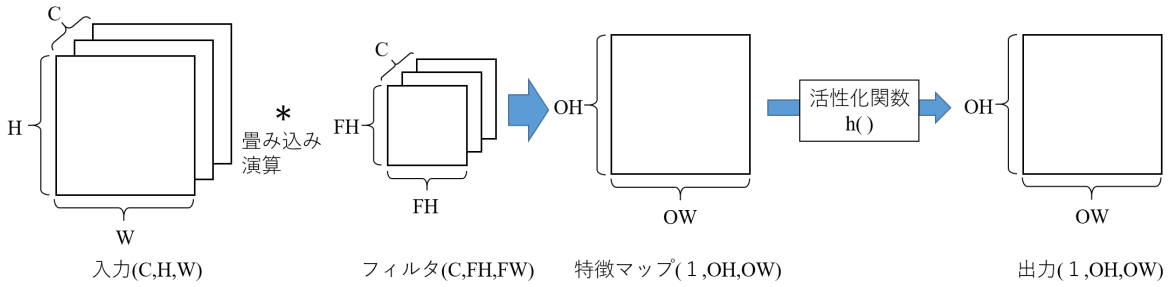

図 2.2　Convolution 層

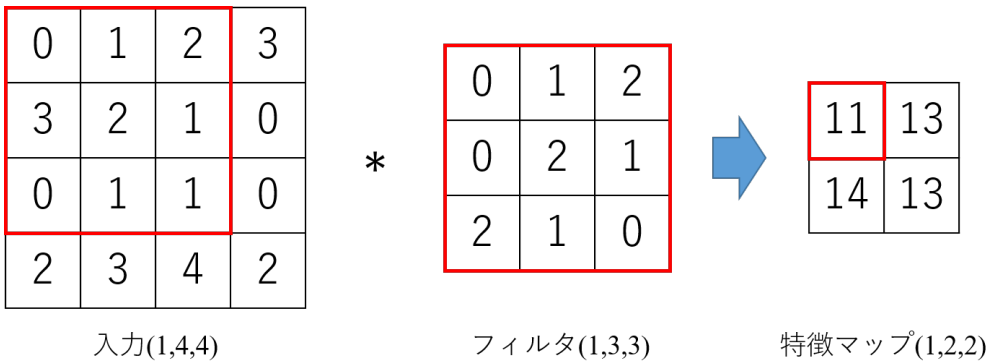

図 2.3　畳み込み処理

2.2.2　Deconvolution 層

Deconvolution 層は、Convolution 層と逆の操作をし、特徴マップの大きさを拡大するものである。つまりアップサンプリングするものである。処理は、図 2.4 のように入力データの各要素間に値が 0 の要素を挿入し拡張し、Convolution 層と同様にフィルタの畳み込み処理を行う。Convolution 層と同様に入力サイズ、フィルタサイズおよびストライドにより特徴マップのサイズが決定する。

用途としては、ニューラルネットワークにおいて、特徴マップから画像生成等を行う際に使用される。

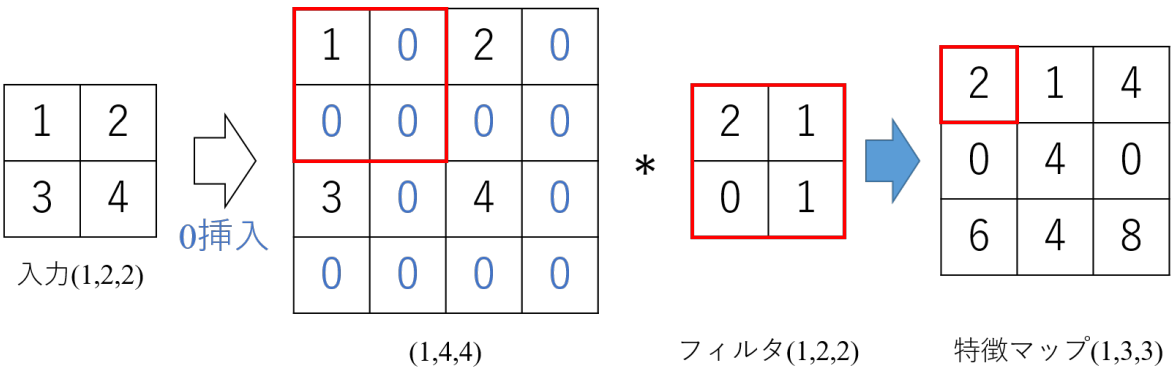

図 2.4　Deconvolution 層

2.2.3　全結合層

全結合層とは、図 2.1 で示した様な一般的なニューラルネットワークであり、隣接する層のすべてのニューロン間が結合されているものである。入力データは 1 次元の行列にする必要があり、出力も 1 次元の行列になる。

CNN では、Convolution 層により空間的特徴を抽出した特徴マップに対して、1 次元行列に変換する整列を行い、全結合層を用いることで、画像などの分類を行う際に使用される。

2.3　GAN（Generative Adversarial Networks）

GAN（Generative Adversarial Networks）[11] は、敵対的生成ネットワークと呼ばれる生成モデルの一種である。GAN では、単一のニューラルネットワークの学習を行うのではなく、生成用ネットワーク (Generator) と識別用ネットワーク (Discriminator) の 2 つの競合するネットワークの学習を行う。Generator では入力データから所望の出力をするように学習を行う。一方で、Discriminator 側では、本物データ (教師データ) と Generator が作った生成データの識別を行うように学習を行う。Generator 側は、Discriminator が誤認識するように画像を生成するように競合させることで、生成精度を向上させる。

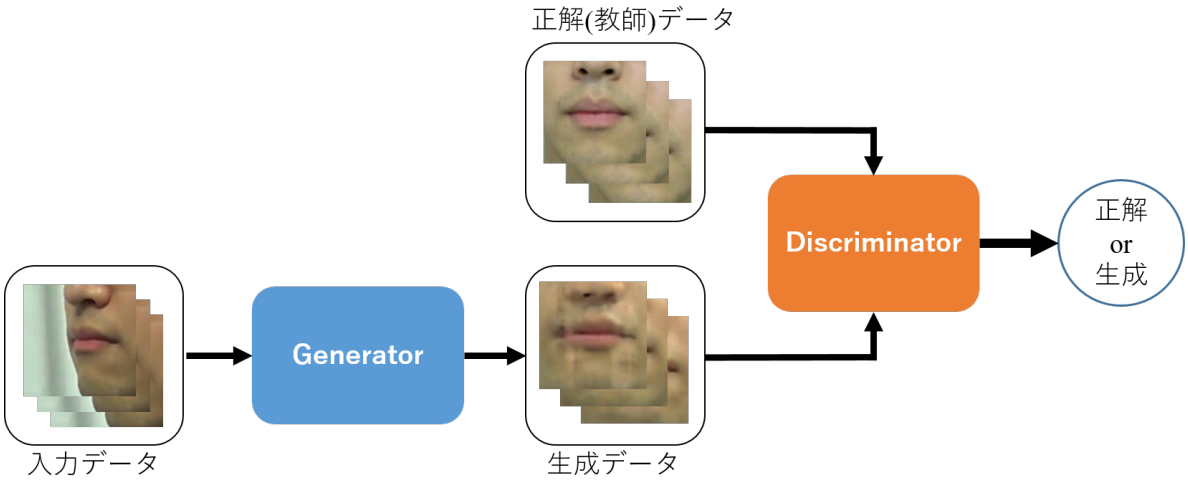

図 2.5　GAN（Generative Adversarial Networks）

第 3 章

提案手法

3.1　正面変換について

本提案手法の目的は、正面視点のみを学習させたリップリーディングのモデルを非正面視点の入力に対応させることである。この目的達成のために、図 3.1 に示すようにリップリーディングの前処理として CNN を用いて入力画像を正面視点へ変換する。

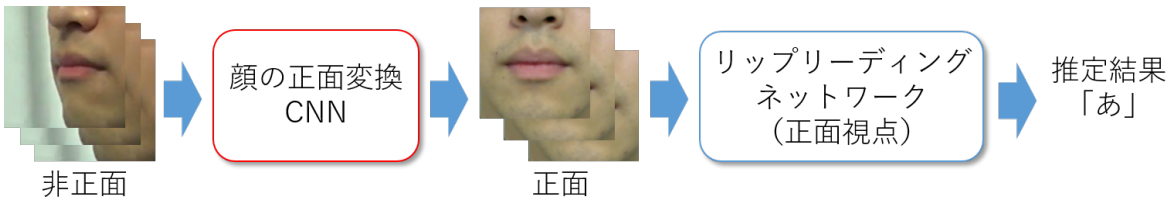

図 3.1　提案手法の流れ

3.2　モデルの構成

本論文で正面変換を行うニューラルネットワークモデルの構成は図 3.2、図 3.3 に示すように、Generator と Discriminator の 2 つのネットワークをもつ GAN の構成をした CNN になっている。このモデル構成は Rui Huang らの顔認証のための正面顔生成ネットワーク [12] の一部を参考にした。モデルの詳細なパラメータは表 3.1、表 3.2、表 3.3 に示す。

Generator は Convolution 層と Deconvolution 層からなるエンコーダデコーダの構成になっており、図中の矢印はスキップ接続を意味している。入力データは、RGB3 チャンネルの画像を左右反転させたものを結合することにより 6 チャンネル画像になっている。これは、入力画像の撮影角度が大きい際に口元の情報が画像の左右片側に偏るため、反転

させたものを結合することで左右どちらにも口元の情報が伝搬するようにした。活性化関数は出力層で tanh 関数、入力側の 4 つの Convolution 層では LeakyReLU 関数、それ以外は ReLU 関数を用いた。また、オプティマイザは Adam を使用した。

Discriminator は Convolution 層と ResBlock で構成されている。活性化関数は出力層のみシグモイド関数、それ以外は ReLU 関数を用い、オプティマイザは Adam を使用した。

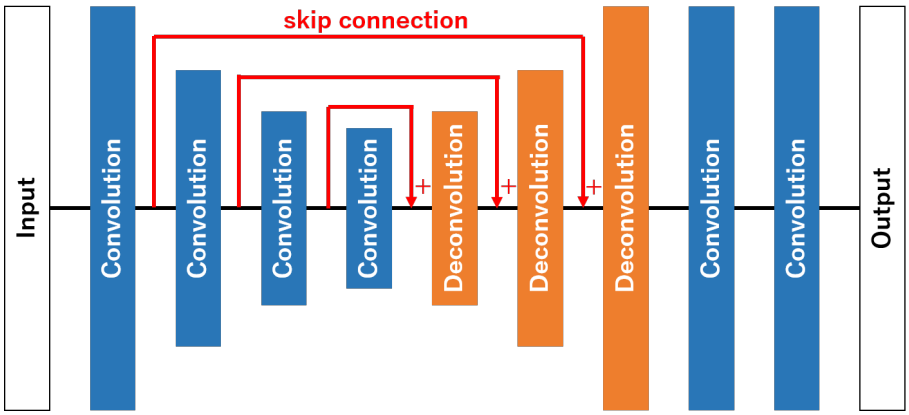

図 3.2　Generator

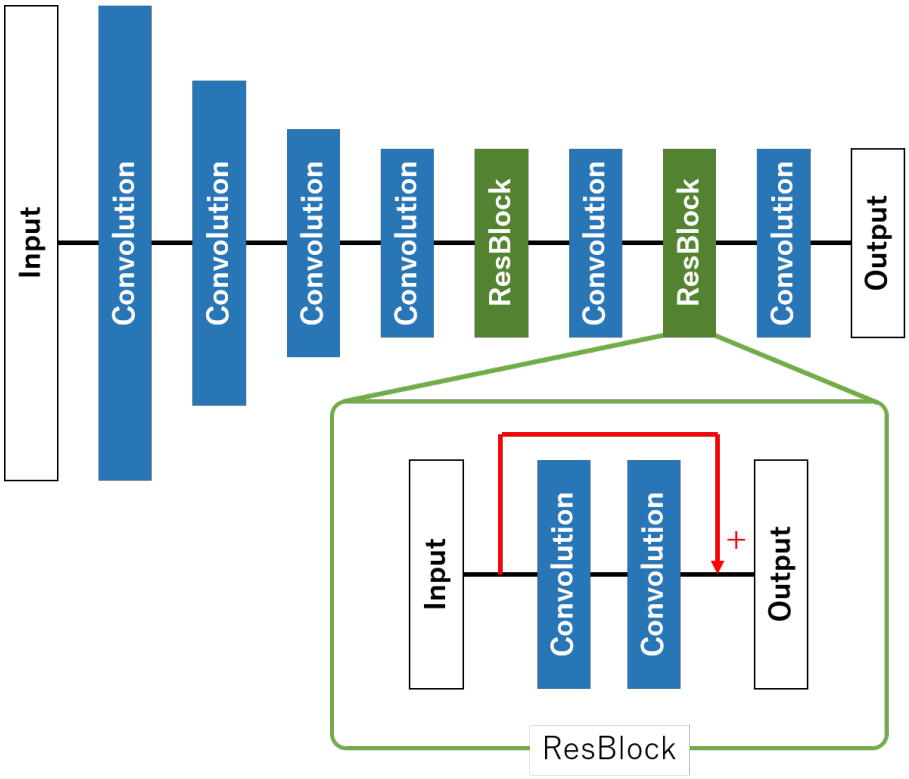

図 3.3　Discriminator

表 3.1　Generator の構成

layer	input	output	filtersize	stride
conv0	h×w×6	h×w×64	3×3	1
conv1	conv0	h/2×w/2×128	3×3	2
conv2	conv1	h/4×w/4×256	3×3	2
conv3	conv2	h/8×w/8×512	3×3	2
deconv0	conv3	h/4×w/4×256	3×3	2
deconv1	deconv0+conv2	h/2×w/2×128	3×3	2
deconv2	deconv1+conv1	h×w×64	3×3	2
conv4	deconv2+conv0	h×w×64	3×3	1
conv5	conv4	h×w×3	3×3	1

表 3.2　Discriminator の構成

layer	input	output	filtersize	stride
conv0	h×w×3	h/2×w/2×64	3×3	2
conv1	conv0	h/4×w/4×128	3×3	2
conv2	conv1	h/8×w/8×256	3×3	2
conv3	conv2	h/16×w/16×512	3×3	2
ResBlock0	conv3	h/16×w/16×512		
conv4	ResBlock0	h/16×w/16×512	3×3	1
ResBlock1	conv4	h/16×w/16×512		
conv5	ResBlock1	h/16×w/16×1	1×1	1

表 3.3　ResBlock の構成

layer	input	output	filtersize	stride
conv0	h×w×c	h×w×c	3×3	1
conv1	conv0	h×w×c	3×3	1

次に、提案モデルの学習に用いる損失関数について述べる。

Generator の損失関数は式 (3.1) のように、3 つの損失の合成から成る。

$$L_{gen} = \lambda_{pixel} L_{pixel} + \lambda_{sym} L_{sym} + \lambda_{adv} L_{adv} \tag{3.1}$$

L_{pixel} はピクセル損失、L_{sym} は対称性損失、L_{adv} は敵対性損失である。λ はそれぞれの重みづけのための係数であり、$\lambda_{pixel} = 1.0$、$\lambda_{sym} = 0.5$、$\lambda_{adv} = 0.5$ に設定した。

チャンネル C、幅 W、高さ H の画像の場合を考える。ピクセル損失 L_{pixel} は、Generator の生成画像と正解画像のピクセル値の平均絶対誤差 (Mean Absolute Error:MAE) を用いる。式 (3.2) で求められ、$I^{pred}_{c,x,y}, I^{gt}_{c,x,y}$ はそれぞれ Generator の生成画像と正解画像のピクセル値を表し、c, x, y は画像におけるピクセルの座標を表す。

$$L_{pixel} = \frac{1}{C \times W \times H} \sum_{c=1}^{C} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| I^{pred}_{c,x,y} - I^{gt}_{c,x,y} \right| \tag{3.2}$$

画像生成系の CNN において平均二乗誤差 (Mean Squared Error : MSE) より MAE の方を損失に用いたが方が視覚的にも複数の定量的指標を用いた場合でも優れているという結果が報告 [10] されているため MAE を採用している。

対称性損失 L_{sym} は、人間の顔の特徴である対象性を考慮するものである。これは、撮影角度が大きいことによる自己オクルージョン問題を軽減させるための損失である [12]。対称性損失は式 (3.3) で求められ、生成画像の左右の対応するピクセル値の MAE である。

$$L_{sym} = \frac{1}{C \times W/2 \times H} \sum_{c=1}^{C} \sum_{x=1}^{W/2} \sum_{y=1}^{H} \left| I^{pred}_{c,x,y} - I^{gt}_{c,W-(x-1),y} \right| \tag{3.3}$$

敵対性損失 L_{adv} は、Discriminator の識別による損失である。Discriminator へ生成画像を入力時の出力がサイズ $W_d \times H_d$ の $F^{sf}_{x,y}$ の場合を考える。敵対性損失は、式 (3.4) により求められ、Discriminator の出力が 0 に近づくほど損失値が小さくなる。

$$L_{adv} = \frac{1}{W_d \times H_d} \sum_{x=1}^{W_d} \sum_{y=1}^{H_d} F^{sf}_{x,y} \tag{3.4}$$

敵対性損失を考慮することにより、生成画像のぼやけが改善される [12]。

一方で Discriminator の損失 L_{dis} は、Discriminator へ正解画像を入力時の出力が $F^{gt}_{x,y}$ としたとき、式 (3.5) で求める。

$$L_{dis} = \frac{1}{W_d \times H_d} \sum_{x=1}^{W_d} \sum_{y=1}^{H_d} \left(F^{sf}_{x,y} - 1 \right) + \frac{1}{W_d \times H_d} \sum_{x=1}^{W_d} \sum_{y=1}^{H_d} F^{gt}_{x,y} \tag{3.5}$$

$$\tag{3.6}$$

Discriminator の出力が生成画像入力時に 0、正解画像入力時に 1 に近づくほど損失値が小さくなる。L_{adv} と L_{dis} の損失により、Generator の生成と Discriminator の識別を競合させる。

以上の損失関数の最小化問題を勾配降下法により解き、誤差逆伝搬法により CNN のパラメータを更新することでネットワークを学習する。

3.3　学習

学習には自作したデータセットを使用した。被験者 15 名が日本語 46 種 (あ, い,···, を, ん) 発音時の口元を 0°、30°、45°、60°、90° から撮影した。被験者 15 名中 10 名を学習用、5 名をテスト用とし、各角度ごとに、学習用で 73,600 データ、テスト用で 36,800 データ作成した。データセットの詳しい作成方法は付録 A に記載する。

提案手法のネットワークを大きく 2 つのパターンで学習を行った。ひとつは、学習用に撮影した 30°、45°、60°、90° の角度ごとにそれぞれ別のネットワークで正面変換の学習を行った。もうひとつは、複数角度からの正面変換をひとつのネットワークで学習した。

3.3.1　角度ごとの正面変換

30°, 45°, 60°, 90° の入力に対して正面視点画像を出力するように各角度ごとにネットワークで学習を行った。学習に用いていない被験者の口元画像の入力に対する、正面視点画像の生成結果は図 3.4、図 3.5、図 3.6、図 3.7 のようになった。

生成結果より、入力画像の角度が大きくなるほど生成画像の精度が落ちているのがわかる。30°、45°、60° において、おおよそ口の形状が読み取れる画像の生成が確認でき、90° になると生成画像が大きく崩れることが確認できた。

現状のモデルでは、90° の入力に対応できないため、本論文では 60° までの入力角度を考慮する。

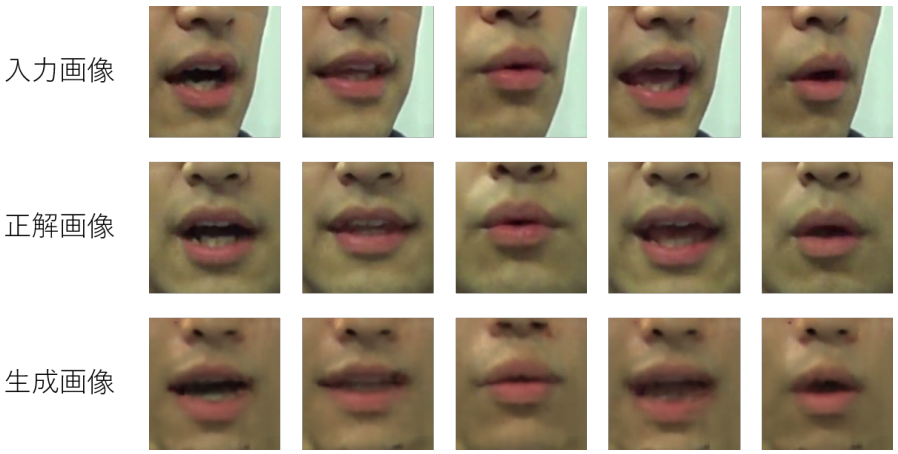

図 3.4　学習角度 30° の生成画像

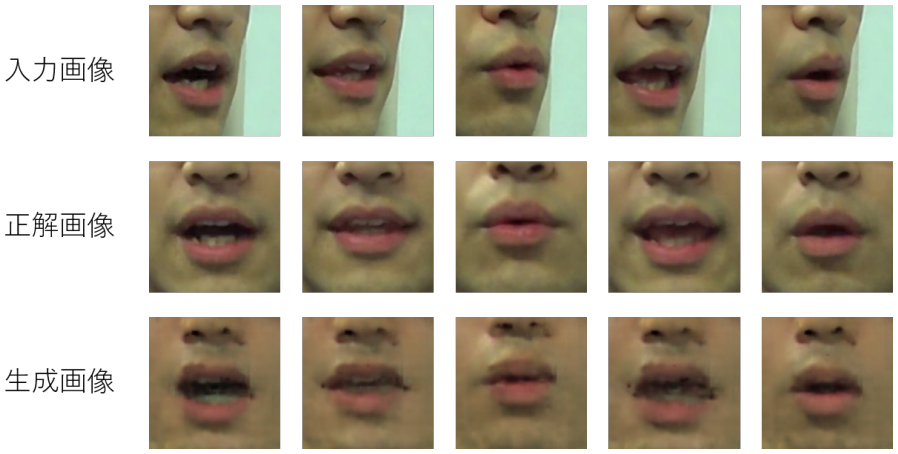

図 3.5　学習角度 45° の生成画像

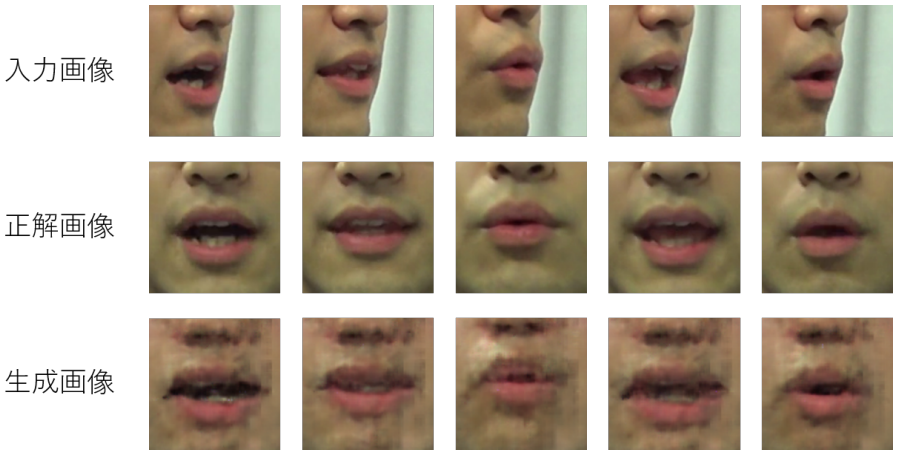

図 3.6　学習角度 60° の生成画像

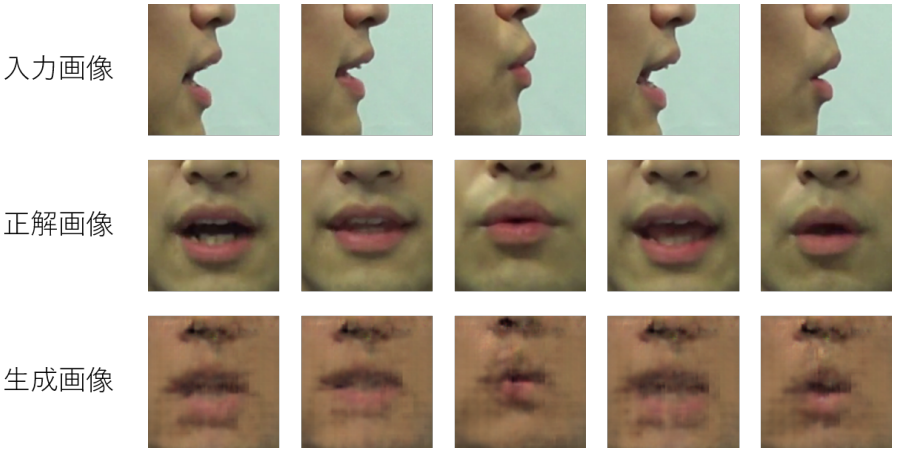

図 3.7　学習角度 90° の生成画像

3.3.2　複数角度の正面変換

3.3.1 節の結果より、60° までの入力を考慮し、0°, 30°, 45°, 60° の入力に対して正面視点画像を出力するようにひとつのネットワークで学習を行った。学習には各角度ごとに 73,600 データの合計 294,400 データを用いた。

複数角度の変換を学習させたネットワークに対して、学習データに用いていない被験者のデータを入力した際の生成画像を図 3.8 に示す。入力角度が大きくなるほど正解画像と生成画像を比較したときの口の形状の精度が落ちる傾向が見られるが、0°, 30°, 45°, 60° のすべてにおいて口として視覚的に認識できる画像が生成できていることが確認できる。

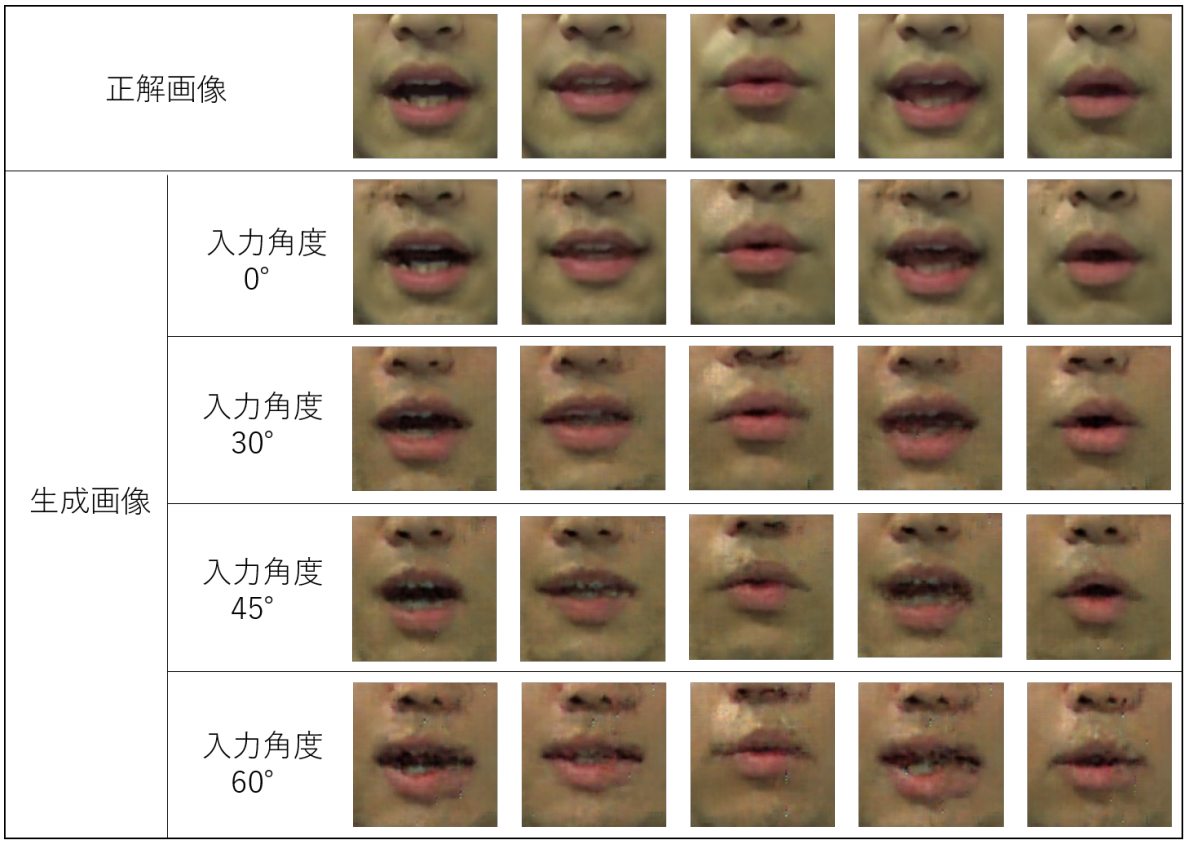

図 3.8　複数角度学習モデルの生成画像

第 4 章

実験

本章では、正面視点のみを考慮したリップリーディングモデルの学習を行い、非正面入力に対して提案手法の適用により認識率に対して有用性があることを示す。

4.1 実験条件

4.1.1 モデル構成 (リップリーディング)

提案手法の有用性を検証するため、図 4.1 に示すリップリーディングのモデルを使用した。このモデルは、Joon Son Chung らのリップリーディングモデル [4] を簡略化したものであり、3 層の Convolution 層と 2 層の全結合層で構成されている [13]。詳しいネットワークのパラメータは表 4.1 に示す。活性化関数は ReLU 関数を用いている。

入力は、120×120 のグレースケール画像を 10 フレーム分を結合し、120×120×10 のデータを用いる。これにより、発話時の時系列情報を学習させる。

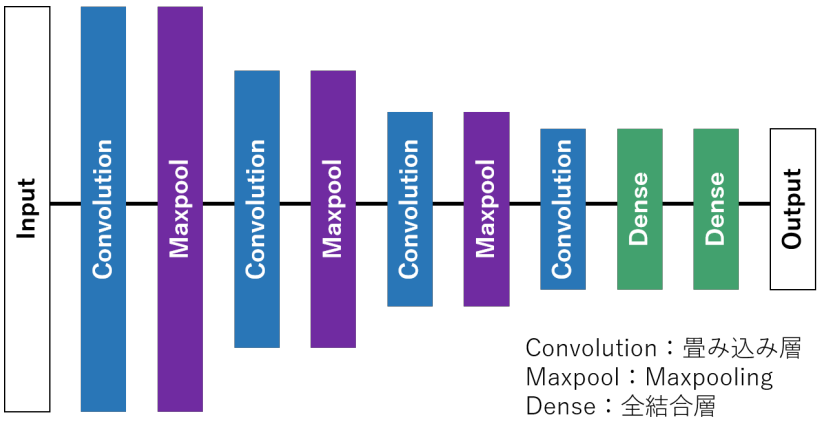

図 4.1　リップリーディングのモデル

表 4.1　リップリーディングのモデル構成

layer	input	output	filtersize	stride
conv0	120×120×10	116×116×128	5×5	1
maxpool0	116×116×128	58×58×128	2×2	2
conv1	58×58×128	54×54×64	5×5	1
maxpool1	54×54×64	27×27×64	2×2	2
conv2	27×27×64	23×23×32	5×5	1
maxpool2	23×23×32	11×11×32	2×2	2
dense0	3872	20		
dense2	20	5		

4.1.2　データセット (リップリーディング)

リップリーディング用のデータセットについて述べる。11 名の被験者に対して、日本語の母音「あ、い、う、え、お」5 種を発音した際の様子を撮影した。撮影は正面変換用 CNN のデータセットと同様に行った。撮影条件は表 4.2 に示す。撮影角度は、正面変換に用いた 0°, 30°, 45°, 60° の他に、間の角度である 15°, 37.5°, 52.5° の撮影も行った。1 音あたり 40[frame] で撮影し、そこから間引いた 10[frame] を 1 音のデータセットとした。

また、データ数を補強するために、口元の切り出しの際に ±10 ピクセルの範囲でランダムに縦・横方向へのシフトと ±5° の範囲でランダムな回転を加えることにより、データ数を 4 倍に増やした。その結果、各角度ごとに、学習用で 5250、テスト用で 3000 データを作成した。ただし、学習には正面角度のみを使用する。

表 4.2　リップリーディング用データ撮影条件

被験者数	11(学習 7 名、テスト 4 名)
撮影角度	0°, 15°, 30°, 37.5°, 45°, 52.5°, 60°
発音の種類	日本語母音 5 種 (あ、い、う、え、お)
各被験者の発音回数	1 音あたり 30 回
フレームレート	30[fps]
1 音あたりのフレーム数	40[frame]
入力データに使用するフレーム数	10[frame]

4.1.3　学習

4.1.1 節のリップリーディングモデルを 4.1.2 節で作成したデータセットで学習を行った。データセットの被験者 11 名のうち 7 名の被験者の正面視点データのみを学習に用いた。データ数 5,250、バッチサイズ 20、オプティマイザは Adam で学習を行った。

4.2　実験結果

4.1.3 節で学習したリップリーディングモデルに対して、0°, 15°, 30°, 37.5°, 45°, 52.5°, 60° のデータセットを未変換で入力した場合と提案手法を用いて正面視点へ変換して入力をした場合の母音 5 種類の分類を行った。

評価指標として、認識率を式 (4.1) で計算した。

$$\text{認識率} = \frac{\text{正しく分類したデータ数}}{\text{総データ数}} \times 100 \quad [\%] \tag{4.1}$$

表 4.3 に、提案手法を用いない場合と用いた場合の母音 5 種類の認識率を示す。ここでは、学習に用いていないテスト用被験者 4 名のもののみ使用した。

また、図 4.2〜4.9 に正解に対して予測結果がどのように分布しているかを示す混同行列を示す。ここでは、提案手法を用いない未変換で分類した結果と 0°, 30°, 45°, 60° を学習した正面変換を適用した分類の結果を示す。

表 4.3　認識率の比較

		未変換	学習角度 [°]			
			30	45	60	0,30,45,60
入力角度 [°]	0	62.4	57.3	60.4	51.5	**62.9**
	15	**58.1**	57.8	56.9	56.0	53.3
	30	39.0	**54.5**	50.0	43.5	52.2
	37.5	40.2	**53.4**	52.7	44.6	50.7
	45	24.1	46.7	50.8	46.2	**53.1**
	52.5	26.0	50.6	**53.6**	50.0	51.8
	60	25.8	42.0	45.6	47.2	**47.9**
平均		37.8	50.1	51.7	47.1	**53.8**

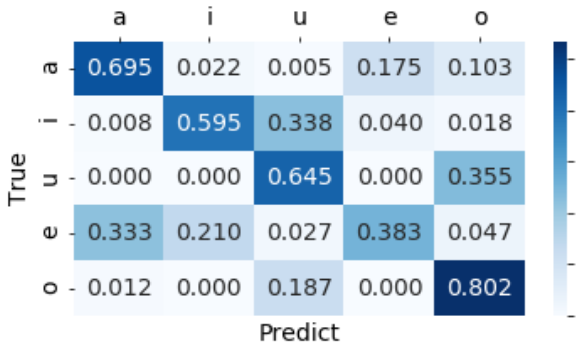

図 4.2　未変換 (入力 0°)

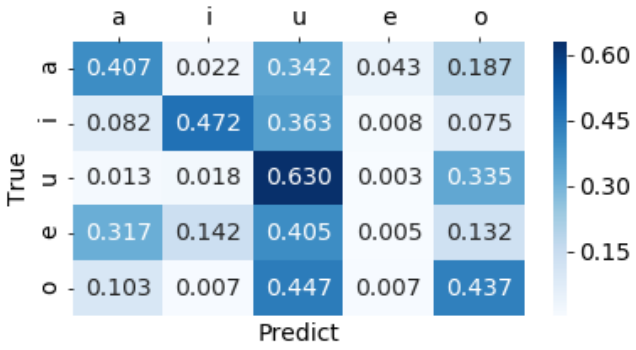

図 4.3　未変換 (入力 30°)

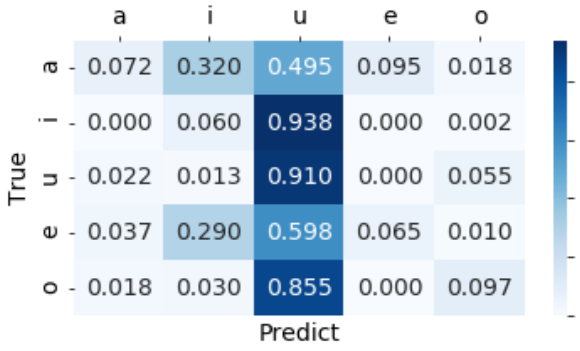

図 4.4　未変換 (入力 45°)

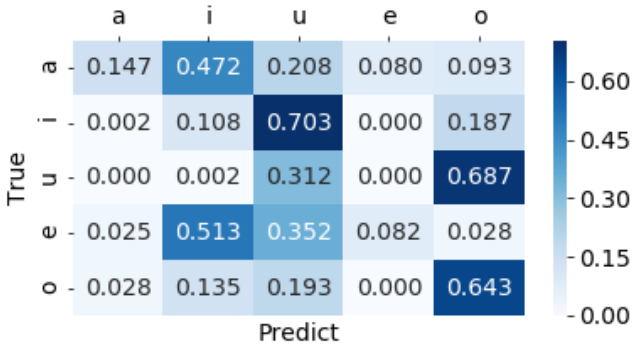

図 4.5　未変換 (入力 60°)

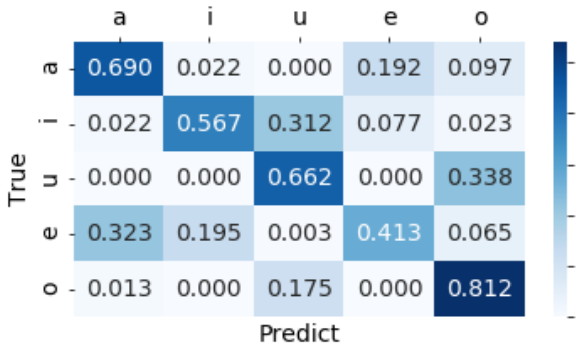

図 4.6　0°, 30°, 45°, 60° 学習 (入力 0°)

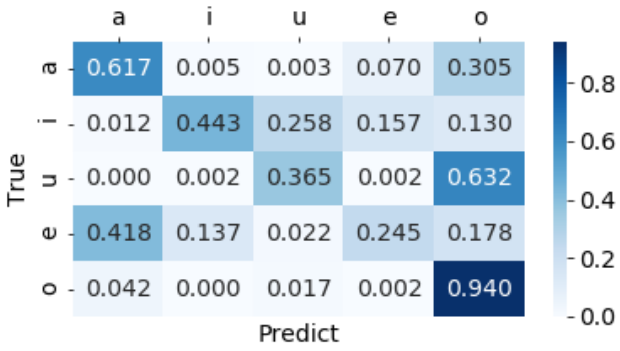

図 4.7　0°, 30°, 45°, 60° 学習 (入力 30°)

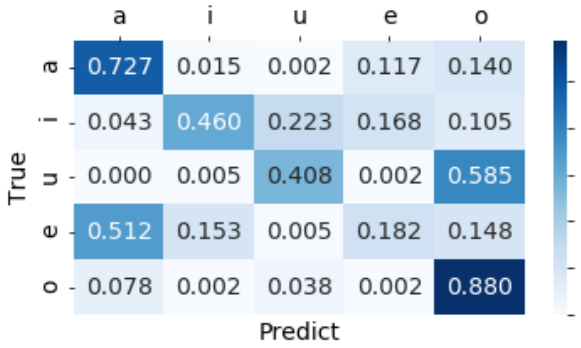

図 4.8　0°, 30°, 45°, 60° 学習 (入力 40°)

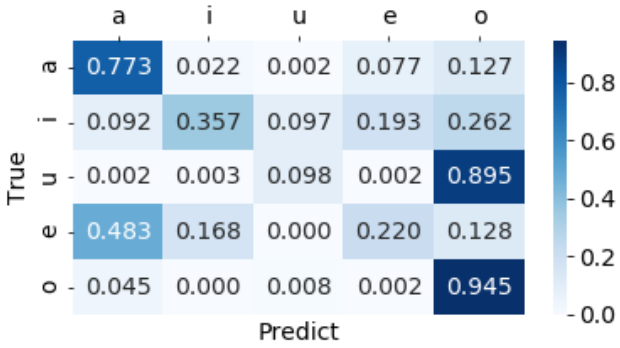

図 4.9　0°, 30°, 45°, 60° 学習 (入力 60°)

4.3　考察

表 4.3 より、正面のみ学習したリップリーディングモデルに対しての認識率は 6 割程度となった。入力角度が大きくなるほど認識率が大きく低下する傾向もみられ、60° で約 25% 程度まで認識率が低下した。図 4.2 より、0° 入力の場合の予測の分布をみると、「え」の認識率が少し低く、「お」の認識率が高いが、完全に予測の分布が偏っている状態ではなく、このモデルで日本語母音の分類ができているといえる。「え」の予測結果が、「あ」と「い」に分類される場合が見られる、これは「え」の発音時の口の開きが大きいと「あ」、「え」の発音時の口の開きが小さいと「い」の形と近くなるため分類が困難であると考えられる。図 4.4 では、完全に予測が「う」に偏り、図 4.5 では予測がばらけており、入力角度がつくと分類できていないことが確認できる。

提案手法を用いた場合の結果について述べる。各角度ごとで正面変換を学習した場合、正面付近 (0°, 15°) の入力では未変換で分類をしたほうが認識精度が高いが、入力角度が大きくなると提案手法を用いることで、認識精度が向上していることが確認できた。

一方で、複数角度で正面変換を学習した場合、正面付近 (0°, 15°) の入力に対しても未

変換と同等の認識率であり、その他の入力角度であっても認識率が 5 割前後であり、提案手法を用いない未変換と比べて認識精度が約 2 割向上したことが確認できた。図 4.6 の複数角度で学習した場合で入力が正面の混同行列をみると、図 4.2 の未変換の場合と同じような予測の分布になっていることから、提案手法の変換を用いても口の形状が保持されていると考えられる。図 4.7〜4.9 より、「う」が「お」に分類されることが多く見られる。これは、「う」と「お」はどちらも口をすぼめるような形で類似しているため、正面変換時に「う」が「お」に近い形状で出力されたのではないかと考える。

認識率の平均を見ると複数角度で学習させたモデルを用いた場合が最も認識率が高いことや、角度ごとにモデルを分けないので入力がどの角度であるかの判断がいらないという点から、一つのモデルで複数角度の正面変換を学習させた方が角度ごとのモデルで学習するよりも良いと考える。

第 5 章

おわりに

5.1 まとめ

本論文では、正面視点のみのデータセットで学習されたリップリーディングモデルを非正面視点の入力に対応させることを目的として、CNN による正面変換を提案した。第 1 章では、リーディングモデルについての研究動向についてと現在の問題点について述べ、本研究目的を示した。第 2 章では、提案手法で用いる CNN ついて基礎理論を述べた。第 3 章では、提案手法である正面変換で用いた CNN のモデルについて、モデルの構成と学習結果について述べた。学習は入力角度ごとに別のネットワークで学習を行うパターンと複数の入力角度をひとつのネットワークで学習を行う 2 パターンを行った。角度ごとの学習モデルの生成画像の結果から本論文では 60° までの入力を考慮することとし、複数角度の学習では、0°, 30°, 45°, 60° の正面変換を学習させ、生成画像で口元画像が生成されていることを確認した。第 4 章では、提案手法の有用性を検証するために、ニューラルネットワークを用いたリーディングモデルを正面視点のみのデータセットで母音 5 種の分類を学習し、0°, 30°, 45°, 60° のデータを未変換でそのまま入力した場合と提案手法を適用して分類結果の比較を行った。角度のある入力に対して、提案手法により正面変換を行うことにより認識精度が向上することが確認できた。また、複数角度を学習させたモデルでは、0° の入力の際にも未変換の場合と比べて認識精度が落ちることはなかった。よって、CNN を用いた正面変換により、正面視点のみ考慮したリップリーディングモデルに非正面入力したときの認識精度向上において有用性が示せた。

5.2 今後の課題

本論文では、日本語母音 5 種のリップリーディングに対して、提案手法により非正面入力に対する認識精度の向上を示した。しかし、現在行われているリップリーディングの

研究では単語や文章レベルでもっと多くの分類を扱っている。従って、母音 5 種のみでなく、もっと複雑なモデルに対しての有用性の検証が必要である。

また、現在の CNN による正面変換の精度も十分とは言えない。原因の一つとしては、データセットの数が少ないことが考えられるため、回転やシフトによるデータ数増強ではなく元データの数を増やす必要がある。その他にも、ネットワークモデルのパラメータに関しても実験的に最適な値を決めることで生成画像の精度が上がる可能性がある。

よって、他モデルでの有用性の検証と正面変換の精度向上が今後の課題といえる。

謝辞

本研究を行うにあたり、ご指導ご鞭撻いただいた、杉田泰則准教授に厚く感謝を申し上げます。また、本論文の審査において貴重な助言をいただいた、本学の岩橋政宏教授ならびに圓道知博准教授に感謝いたします。さらに、被験者の協力と日々の研究生活でお世話になりました信号処理応用研究室の皆様に心から謝意を表します。

令和 2 年 2 月 7 日

参考文献

[1] N.L. Hesselmann, “Structural analysis of lip-contours for isolated spoken vowels using fourier descriptors,” Speech Communication, Vol.2, pp.327-340 1983.

[2] 高橋　昌平, 大谷　淳, “複数画像特徴を用いた読唇システム-オプティカルフロー特徴・形状特徴・離散コサイン変換特徴の統合の検討-,” 情報処理学会研究報告. CVIM, pp.1-7, 2014.

[3] 松岡清利, 古谷忠義, 黒須 顕二, “画像処理による読唇の試み-母音口形の識別およびそれに基づく単語認識-,” 計測自動制御学会論文集,Vol.22, pp.1-7, 1986.

[4] Joon Son Chung, Andrew W. Senior, Oriol Vinyals , Andrew Zisserman, “Lip Reading Sentences in the Wild,” IEEE CVPR, pp.3444-3453, 2017.

[5] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas, “Lipnet:Sentence-level lipreading,” ICLR, 2017.

[6] Noda K., Yamaguchi Y., Nakadai K., “Audio-visual speech recognition using deep learning,” Applied Intelligence, Vol.42, pp.722-737, 2015.

[7] Joon Son Chung, Andrew Zisserman, “Lip Reading in Profile,” British Machine Vision Conference, 2017.

[8] Ziheng Zhou and Guoying Zhao and Xiaopeng Hong and Matti Pietikainen, “A review of recent advances in visual speech decoding,” Image Vision Comput, Vol.32, pp.590-605, 2014.

[9] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, “ImageNet Classi
cation with Deep Convolutional Neural Networks,” NIPS’12: Proceedings of the 25th International Conference on Neural Information Processing Systems2,Vol.1, pp.1097-1105, 2012.

[10] Hang Zhao, Orazio Gallo, Iuri Frosio, Jan Kautz,, “Loss Functions for Image Restoration With Neural Networks,” IEEE Transactions on Computational Imaging, Vol.3, pp.47-57, 2017.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” NIPS, 2014.

[12] R. Huang, S. Zhang, T. Li and R. He, “Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis,” 2017 IEEE International Conference on Computer Vision (ICCV), pp.2458-2467, 2017.

[13] 阿部　匠, “顔回転と機械学習を用いた複数視点リップリーディングに関する研究,” 長岡技術科学大学大学院工学研究科修士論文, 2019.

付録

A　データセット

本論文で正面変換に用いるデータセットは、自作したデータセットを用いた。ここでは、その作成方法を述べる。

斜め視点の口元画像から正面視点の画像への変換を学習させるために、15 名の被験者が日本語の五十音 46 種を発音した際の様子を 5 台のカメラで撮影した。撮影条件は表 A.1 にまとめる。撮影には、SONY のデジタルビデオカメラレコーダ HDR-AS50 を用いた。図 A.1 に示す様に、椅子と頭部固定器具の周りに 5 台のカメラを設置して撮影を行う。頭部固定器具は撮影時に被験者の頭部の位置が移動しないように 2 本の棒材で頭部を挟み、固頭部を定するために用いる。撮影する角度は 0°(正面), 30°, 45°, 60°, 90° の 5 方向である。5 方向から撮影した例として図 A.2 に動画の 1 フレームを示す。

表 A.1　正面変換学習用データの撮影条件

撮影環境	室内
被験者数	15 名 (学習 10 名、テスト 5 名)
撮影角度	0°(正面),30°, 45°, 60°, 90°
発音の種類	日本語 46 種 (あ, い,· · · , を, ん)
各被験者の発音回数	1 音あたり 2 回
各音の発音間隔	40[frame] ごと
フレームレート	30[fps]

カメラ等の配置

頭部固定時の様子

図 A.1　撮影環境

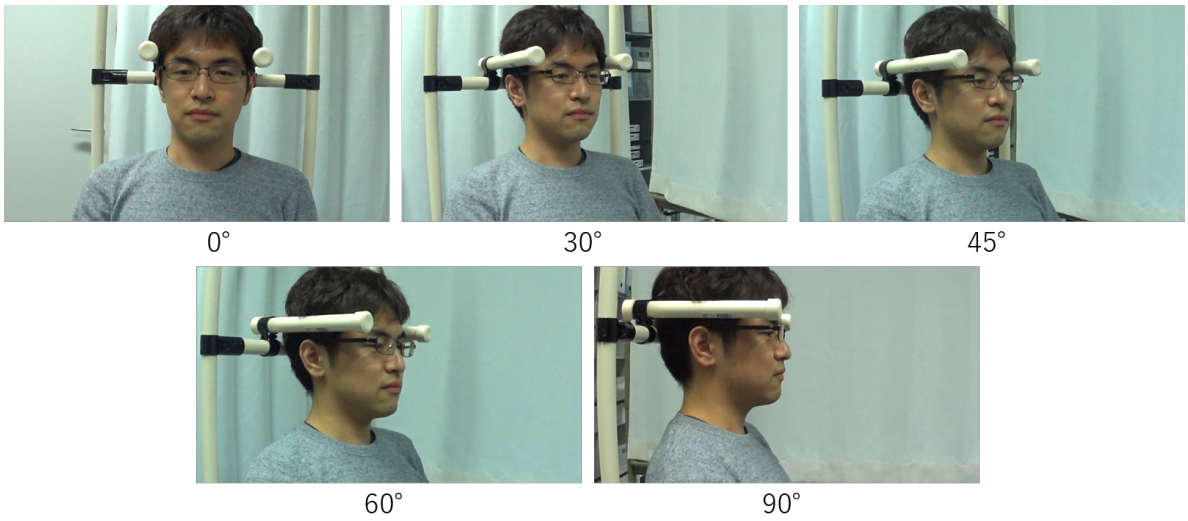

図 A.2　5 方向から撮影した様子

撮影した動画の各フレームから図 A.3 のように口元を正方形で切り出し、被験者 15 名のうち、10 名を学習データ、5 名をテストデータとして使用した。データセットの仕様は A.2 に示す。

また、本論文ではデータ数が十分でないため、口元の切り出しの際に ±10 ピクセルの範囲でランダムに縦・横方向へのシフトと ±5° の範囲でランダムな回転を加えることにより、データ数を 4 倍に増やした。

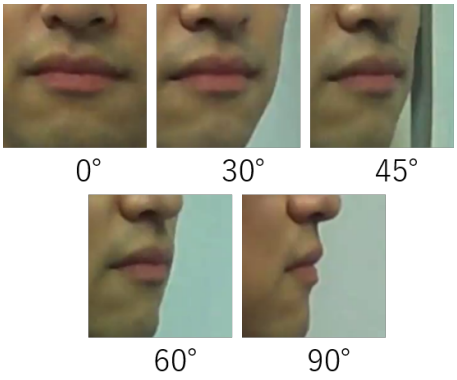

図 A.3　口元を切り出した画像

表 A.2　データセットの仕様

切り出し位置	口元
入力サイズ	3 × 96 × 96
学習データ数 (各角度ごと)	73600
テストデータ数 (各角度ごと)	36800