卒 業 論 文 概 要 書

Summary of Bachelor's Thesis

Date of submission: 02/06/2019

学科名 Department	情報通信	氏名 Name	浅見莉絵子	指 導 教 員 Advisor	渡辺 裕 ㊞
研究指導 Research guidance	オーディオビジュアル 情報処理研究	学籍番号 Student ID number	1w153004-1 CD		
研究題目 Title	機械学習による日本語話者の自動読唇 Lip Reading for Japanese Speaker by Machine Learning				

1. まえがき

近年, 動画配信サービスにおいて自動字幕機能が採用され始めている. しかし, 同機能は音声認識技術によってのみ構成されており, 複数人による同時発話や雑音などの影響を受けやすい. そのため, 生成された字幕が実際の発話内容と異なることが多い. そこで, 音声認識技術に加えて、画像認識技術を用いて発話内容を推定することで, 自動字幕機能の精度向上につながると考えられる. さらに, 自動読唇を実現することで, 聴覚障害者とのコミュニケーション支援や, 音声 AI アシスタントの精度向上にも利用可能である. 自動読唇の研究はあらゆる言語でなされている. しかし日本語は母音の数が少なく, 大まかな口の形から発話内容を推測することが難しい. そのため, 日本語話者に対する自動読唇の研究は発展途上にある. そこで本研究では, 英語の文章レベルでの自動読唇を実現している LipNet の日本語発話シーンへの適用を検討する.

2. 日本語読唇の従来手法

日本語の自動読唇手法は, 口形ベースの手法と単語ベースの手法に大別される.

口形ベースの手法では, 発話した際の口形変化を捉え, 口形順から発話された単語を認識する. 日本語の音素を発話する際の口の形をコード化し, コードの組み合わせによって単語を表現する. 認識の際には, 認識したい発話シーンの口形をコード化し, あらかじめ作成した単語のコードとの類似度を求める. この手法では, 認識したい単語すべての口形コードを作成する必要がある. また, 同じ口形順序コードを持つ単語の識別が難しい.

単語ベースの手法は, 単語の発話シーンを事前に撮影し, 特徴量などを用いて発話された単語を認識する手法である. 単語ベースの手法については様々な研究がなされており, そのうちの一つに Active Appearance Model (AAM) を用いた手法がある. 駒井らは, AAM によって唇領域の特徴点を抽出し, AAM によって得られる情報から, 発話した音素を予測する[1]. AAM による音素予測結果は, 母音の正解精度は 67.81%, 子音の正解精度は 11.85%である[1]. また, Convolutional Neural Network (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認識精度は 50.9%である[2]. しかし, どちらの手法も認識精度が低く, 画像のみからの完全な発話内容認識には至っていない.

3. LipNet

本研究は LipNet を日本語のデータセットに適用し, 日本語の自動読唇を行う.

LipNet は, 英語話者の自動読唇システムであり, 英語発話に対する認識精度は 93.4%である. 同じ発話動画をプロの読唇術者が読み取った場合の認識精度は 52.3%であり, 人間の認識精度をはるかに上回っている. LipNet は, プロの読唇術者が長い言葉ほど正確に読み取れるという特徴を応用している. 従来の単語レベルの検出ではなく, 文章レベルの学習を行うことで読唇精度を向上させている. 図 1 に LipNet 構成を示す. T フレームのシーケンスを入力し, 時空間の畳み込みニューラルネットワークである Spatiotemporal Convolutional Neural Networks (STCNN) の三つのレイヤで処理される[3]. STCNN によって抽出された特徴は, Recurrent Neural Network (RNN) の一種である Gated Recurrent Unit (GRU) によって双方向に処理される. GRU 出力のタイムステップには線形変換が適用され, Softmax が適用される. LipNet モデルは Connectionist

Temporal Classification (CTC) で訓練されている.

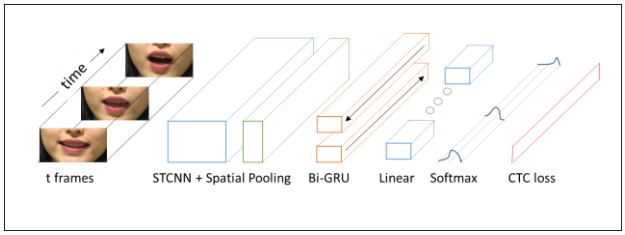

図 1 LIpNet architecture [3]

4. データ作成

日本語には母音が五つしかなく，母音が 30 個近くある英語に比べ, 口形の違いによって見分けることが難しい. また, 英語には口形変化がある子音も存在し, 口形変化において日本語は英語と大きく異なる.

日本語の読唇用データセットは少なく, 公開されている単語レベルのデータセットでは，齊藤らの Speech Scene database by Smart Device (SSSD)が存在する[4][5][6]. しかし，LipNet は文章レベルの学習を前提に作られているため，SSSD では認識がうまくいかない. そこで本研究では，日本語の文章レベルのデータセットを作成した. 1 発話につき五つの単語からなる文を発話したものを 360 発話分用意し，データセットとした. データセットは，SSSD に含まれる単語を中心に作成したもの (データセット A) と，「p」,「b」,「m」を子音にもつ単語を多く含んだもの (データセット B) の二種類を作成した.

5. 実験

作成したデータをトレーニングデータとし，LipNet 構造に適用した. テストデータには, データセット A, B それぞれに使用した文の発話動画を新たに撮影したものを使った. 本実験では，150 エポックで学習を終了させた. データセット A を適用した結果は，Word Error Rate (WER) が 81.18%, Character Error Rate (CER) が 68.73%であった. データセット B を適用した結果は，WER が 65.07%, CER が 60.67%であった.

学習データが少なく, どちらも認識精度は低いが, データセット B を用いた方が認識精度が上がることが確認された.

6. むすび

本研究では, LipNet を用いた日本語話者の自動読唇手法について検討した. 日本語のデータセットを作成し, LipNet 構造に適用したが, 英語のデータセットを使用した場合に比べ, 高い認識精度を得ることはできなかった. この原因としては，トレーニングデータの数が少ないことが挙げられる. また，トレーニングデータを作る際には, 日本語の音素すべてを学習させるために, 音素バランスを考えたデータセットの作成が必要である. さらに，日本語は唇の動きのみでの発話内容の認識が難しい言語であるため, 口周辺の動きや表情も学習に導入することで，認識精度の向上を目指す.

参考文献

[1] 駒井，宮本，滝口，有木, “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”, 画像の認識シンポジウム (MIRU2010), Vol.109, No.376, pp.357-362, July 2010

[2] 柿原，滝口，有木，三谷，大森，中園, “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, 1-P-35, Mar. 2015

[3] Y. Assel, B. Shillingford, S. Whitesaon, and N. de Freitas, “LiPNeT: End-to-End Sentence-Level Lipreading”, https://arxiv.org/abs/1611.01599, Dec. 2016.

[4] 齊藤, 窪川: “SSSD：スマートデバイスを用いた読唇技術向け日本語データベース”，信学技報, PRMU2017-199 Mar. 2018

[5] 窪川, 齊藤: “SSSD を用いた深層学習による読唇精度に関する検討”，第 21 回 画像の認識・理解シンポジウム（MIRU2018）,PS2-39, Aug. 2018

[6] T. Saitoh, and M. Kubokawa: “SSSD: Speech Scene Database by Smart Device for Visual Speech Recognition”, Proc. of ICPR2018, pp.3228-3232, Aug. 2018.

2018 年度　卒業論文

機械学習による日本語話者の自動読唇

Automatic Lip Reading for Japanese Speaker by Machine Learning

指導教員　渡辺 裕 教授

早稲田大学 基幹理工学部 情報通信学科

1W153004-1

浅見 莉絵子

目次

第 1 章　序論

1.1 研究背景

近年，YouTube などの動画配信サービスが広く利用されている．動画を視聴する際，騒音環境下などでは字幕を表示させて視聴する場合がある．また若者を中心に，音楽やドラマを視聴する際に歌詞やセリフを表示させることが広まっており，音楽やドラマなどへの共感性を高めている．このようなユーザ側の変化に対応するために，動画配信サービスにおいて自動字幕機能が採用され始めている．しかし，同機能は音声認識技術によってのみ構成されており，生成された字幕が実際の発話内容と異なることが多い．さらに，複数人による同時発話や雑音などの影響を受けやすいという問題もある．また，Siri などの音声 AI アシスタントが発達しているが，外出時の使用などの雑音環境下では認識性能が低下することがある[1].

人間は視覚情報も利用して会話をしている．そのため音声認識技術に加えて、画像認識技術を用いて発話内容を推定することで，自動字幕機能や音声 AI アシスタントの精度向上につながると考えられる．さらに，自動読唇が実現されれば，監視カメラの映像など，音声が聞き取りにくい場合にも発話内容の推測が可能となり，犯罪捜査や防止などにも有効であると考えられる[2].

また，聴覚障害者のコミュニケーション手段の一つに「読唇」が存在するが，聞き手に読唇能力が必要となる．一般に，読唇能力の習得には二つの方法がある．一つ目は，人が対面に立ち学習する方法である．これは効果的な学習方法であるが，話し手が不在の場合には学習できず，話し手も読唇技能を持っている必要がある[3]．二つ目は，ビデオ映像を利用した学習である．これは 1 人でも学習できるが，ビデオ映像にない語句の読唇を習得することができない．どちらの方法にも短所があり読唇能力の習得は困難である．しかし，読唇能力を習得することで多くの人とのコミュニケーションを円滑にすることができる[3]．そのため，自動読唇が実現されれば，聴覚障害者との手軽で円滑なコミュニケーションが可能になる．

1.2 目的

画像認識技術のみによって発話内容を認識することができれば，自動字幕や音声 AI アシスタントの精度向上につながる．また，聴覚障害者のコミュニケーション支援に役立てることも可能である．

自動読唇の研究は英語を始めとしてあらゆる言語でなされている．英語では，LipNet と呼ばれる自動読唇システムが研究されている．LipNet の認識精度は 93.4%を実現しており，同じ文章をプロの読唇術者が読み取った場合の認識精度である 52.3%をはるかに上回っている．一方で，日本語は母音の数が少なく，大まかな口の形から発話内容を推測することが難しい．そのため，日本語では自動読唇の研究は発展途上にある．そこで本研究では，LipNet を用いた日本語話者の自動読唇手法について検討する．

1.3 関連研究

関連研究として，柿原らのマルチモーダル音声認識の研究が挙げられる[2]．マルチモーダル音声認識とは，画像処理技術と音声処理技術の両方を用いて発話内容を推測するものである．柿原らは，口話をコミュニケーション手段とする重度難聴者を対象としている．この手法では，顔モデルを Point Distribution Model (PDM) で表現し，Constrained Local Model (CLM) によって唇領域を抽出する．CLM は，PDM と，濃淡パターンを表すアピアランスから作られた特徴点検出器によって構成されている．唇領域を抽出したのち，メルマップ化を行った音声とともにボトルネック構造の Convolutional Neural Network (CNN) に入力してボトルネック特徴量を抽出する．抽出された特徴量を隠れマルコフモデル (HMM) の入力とし，マルチモーダル音声認識を実現する[2]．雑音環境下においてマルチモーダル音声認識を行った結果，音声のみでの認識に比べ，認識精度が 18.1 ポイント上昇している．この結果より，音声認識技術に画像認識技術を組み合わせることで，発話内容の認識精度が改善されることが分かる．

1.4 論文の構成

本論文の構成は以下のとおりである．

第 1 章は，本章であり，本研究の背景，目的，関連研究について述べる．
第 2 章は，日本語読唇の従来手法について述べる．
第 3 章は，英語話者の自動読唇システムである LipNet について述べる．
第 4 章は，データ作成について述べる．
第 5 章は，実験とその結果，及び考察について述べる．
第 6 章は，本研究の結論と今後の課題について述べる．

第 2 章　日本語読唇の従来手法

2.1 まえがき

日本語の自動読唇の従来手法は，大きく分けて口形ベースの手法と単語ベースの手法の二つに分類される．本章では，それぞれの手法について詳しく述べる．

2.2 口形ベースの手法

本節では，日本語の自動読唇における，口形ベースの手法について述べる．宮崎らは，プロの読唇術者が口形の変化に基づいて読唇していることを利用し，発話した際の口形変化を捉え，口形順から発話された単語を認識する手法を提案している[4]．口形変化に基づく認識手法は，「口形ベース手法」と呼ばれている．口形ベース手法では，日本語の母音を発話する際の口形と口を閉じている時の口形 (閉唇口形) の 6 口形を「基本口形」と定義し，発話語を形成する基本口形を記号列で表現する．この記号列を「口形順序コード」と呼ぶ．表 2.2.1 に「あ」から「を」までの日本語 45 音の口形コードを示す．

表 2.1　日本語 45 音の口形コード表[5]

あ	か	さ	た	な	は	ま	や	ら	わ
−A	−A	IA	IA	IA	−A	XA	IA	IA	UA
い	き	し	ち	に	ひ	み		り	
−I	−I	−I	−I	−I	−I	XI		−I	
う	く	す	つ	ぬ	ふ	む	ゆ	る	
−U	−U	−U	−U	−U	−U	XU	−U	−U	
え	け	せ	て	ね	へ	め		れ	
−E	−E	IE	IE	IE	−E	XE		IE	
お	こ	そ	と	の	ほ	も	よ	ろ	を
−O	−O	UO	UO	UO	−O	XO	UO	UO	UO

表 2.1 に示す以外に，濁音や半濁音，拗音のコードも存在する[6]．口形順序コードは，表 2.1 の口形コードを用いて，単語単位で作成される．例えば，「なごや」の口形順序コードは，[IA-OIA] となる[5]．表 2.1 において，それぞれの音素を表す口形コードのうち一つ目の口形を初口形，二つ目の口形を終口形と呼ぶ．単語によっては，一文字目の終口形と二文字目の初口形が同じになる単語も存在し，同じになるときは口形順序コードが省略される．「あ

した」の口形順序コードを，表 2.1 を基に作成すると，[-A-IIA]となるが，[I] が連続しているところは口形が変化しないため，実際の口形順序コードは [-A-I-A] となる．

発話シーンの認識では，基本口形が形成されている区間を検出し，その区間の基本口形の類似度を特徴パラメタとして利用する[4]．基本口形が形成されている区間の検出には，あらかじめ用意した話者の基本口形を用いてフレーム単位でマッチングを行う．マッチングを行う際，口唇領域のオプティカルフローも計測し，口唇領域の移動距離を計測する．移動距離が大きいフレームでは口唇が大きく動いているため，口形変化があると推測できる．また，移動距離の小さい区間では，基本口形を形成していると推測できる．移動距離の大きいフレームと移動距離の小さいフレームを分類し，移動距離の大きいフレームを除いた区間を基本口形形成期間としている[4].

単語認識では，初口形が形成されている区間と終口形が形成されている区間に順序をつけ，i 番目の口形形成区間における特徴パラメタを定義する．認識対象単語に同操作を行い，特徴パラメタの類似度によって単語を認識する．認識の際，違う単語においても特徴パラメタの類似度が同じになる場合があるため，単語の長さも考慮している．

宮崎らは，47 都道府県名を認識対象単語とし，ハイスピードカメラを用いて発話映像を撮影した[1]. 47 都道府県では, 東京と京都以外の都道府県は異なる口形順序コードをもつ. この結果，47 都道府県中 36 の都道府県で認識に成功しており，認識率は 76.60%である．しかしながら，この手法では認識したい単語一つ一つに口形順序コードによるパラメタを定義する必要がある．また，同じ口形順序コードをもち，長さが等しい単語については識別が難しい．

2.3 単語ベースの手法

本節では，日本語の自動読唇における単語ベースの手法について述べる．単語の発話シーンを事前に撮影し，特徴量などを用いて発話された単語を認識する手法を単語ベースの手法と呼ぶ．単語ベースの手法については様々な研究がなされており，本節では三つの手法について述べる．

一つ目は，駒井らの，Active Appearance Model (AAM) を用いた手法である[1]．AAM は，発話者の頭部を固定する必要がなく，発話者の位置に関わらず唇を抽出することができる．まず，AdaBoost 法によって顔領域を抽出する．抽出した顔領域に対して AAM を適用し，唇領域を抽出する. 唇領域を抽出する際, あらかじめ学習した顔全体の AAM と唇領域の AAM を用いることにより，唇領域の位置特定の精度が上がる．また，唇領域に AAM を適用する時，AAM の combined パラメタを特徴量として抽出する．combined パラメタには特徴点の形状と輝度値の情報が含まれており，この二つの情報を用いて隠れマルコフモデル (HMM) を作成する．駒井らは，ATR 音素バランス単語を用い，2160 の単語発話映像を用いて学習を行った．学習に使用していない未知データに対して AAM による音素解析を行った結果, 母音の音素正解精度は 67.81%，子音の音素正解精度は 11.85%であった．音素解析の結果より，画像特徴量のみからの子音の音素識別は難しいことが分かる．特徴量を用いた手法は多く存在し，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を用いた手法などが

ある[7].

二つ目は，義平らによる，フーリエ記述子を用いた手法である[8]．彼らは，先天的に聴覚障害を持ち読唇の訓練を受けた学生に実際に読唇させ，読唇方法や注視点について調査した．学生は，読唇の際には口元に着目し，のどや頬はほとんど見ていなかった．さらに，知識や経験によって情報を補い，発話内容を推測していた[8]．義平らは学生の読唇方法を基に，口の輪郭形状と舌の動きを捉えることで，発話内容を推測する手法について提案している．フーリエ記述子は唇の輪郭形状取得に用いている．「あ」から「ん」までの 46 音の発話動画において唇の輪郭部分の座標値を取得し，フーリエ記述子を算出する．フーリエ記述子を算出する際には，それぞれの音素の発音前も算出し，発音途中との差を求める．20 名の発話映像に対して求めたフーリエ記述子の値の平均値をとり，クラスター分析することで，母音の識別を行っている．子音では，フーリエ記述子の時系列解析を行うことで，識別を試みている．義平らは，「あ」，「さ」，「ま」，「わ」の変化を調べ，特徴的な変化が現れることを発見した．子音の識別については一部の音素での実験のみであり，すべての子音に適用が可能であるとの判断ができない．また，この段階で舌の動きは取り入れられていない．

三つ目は，パリアスカらの深層学習を用いた手法である[9][10]．彼らは，Convolutional Neural Network (CNN) によって学習させている．学習データは，9 単語×50 発話分の発話動画に処理を施したものを使用している．発話動画は，口周りのみを撮影したものを使用し，学習効率を上げるためにグレースケールに変換している．発話動画をフレーム分割し，単語発話区間のフレームを 10 枚連結させて 1 枚の画像とする．作成した 1 枚の画像を学習データとしている．9 個の単語は，三つのグループに別れており，それぞれのグループでは一つの単語を基準に，母音と子音を変更した単語を使用している．CNN の実装には，C++によって実装された深層学習向けのフレームワークである Caffe を利用している[9]. 学習モデルは分類器として作成されており，入力データに対して最も近いと思われるものが出力される．テストはグループごとに行っている．テスト用動画も唇のみを撮影し，学習データと同じ処理を施して 1 枚の画像を作成する．各単語 20 枚のテスト画像を用意し，認識を行った結果は，47.22%であった．この手法では，動画撮影の際に唇のみを撮影する必要があり，一般の動画などへの使用は難しい．そのため，顔全体が写っている動画から唇のみを抽出する技術も合わせる必要があると考えられる．また，認識できるのは学習した単語のみであるため，実用化するには多くの単語の学習データが必要になると考えられる．

2.4 むすび

日本語の自動読唇の手法は口形ベースと単語ベースの二つの方法に大別される．さらに単語ベースの手法については様々な手法が研究されている．本章では，それぞれの手法について詳しく述べた．

第 3 章 LipNet

3.1 まえがき

英語の自動読唇システムである LipNet は，93.4%の検出精度を実現しており，同じ文章をプロの読唇術者が読み取った場合の検出精度である 52.3%を大きく上回っている．本章では，LipNet について詳しく述べる．

3.2 LipNet の概要

これまでの自動読唇の研究では，音素や単語の認識を行うものがほとんどであった．しかし Y．Assael らは，プロの読唇術者は長い単語の方が短い単語よりも読唇精度が上がるということを利用し，文章レベルでの認識を行うことで認識精度の向上を図った[11]．文章レベルの認識を実現することで，文章を予測する前に発話動画を単語ごとに分割する必要性がなくなる利点がある．

LipNet では，トレーニングデータとして文章レベルのデータセットである GRID コーパスを利用している．GRID コーパスは，[command + color + preposition + letter + digit + adverb]というシンプルな文法で作られた文章 1000 文×34 人の発話映像で構成されている．また，それぞれの発話映像に対応したアライメントが付加されている．破損している動画もあるため，使用可能なデータは 32746 発話分である．また，各話者からランダムに選ばれた 225 個の発話シーンが評価に使用される．すべての動画は 3 秒間で，フレームレートは 25fps である．動画は，DLib の顔検出器と，68 個のランドマークを持つ iBug 顔ランドマーク予測子[12]をオンライン Kalman フィルタと組み合わせて処理している．これらのランドマークを使用してアフィン変換を適用し，サイズが 100x50 ピクセルの唇画像を抽出する．動画および画像データは RGB チャネルを標準としている．

トレーニングデータにない話者(Unseen Speakers)の動画を使ってテストした結果は，Character Error Rate (CER) が 6.4%, Word Error Rate (WER) が 11.4%となった. Unseen Speakers の学習では，178epoch で学習を終了している．Unseen Speakers の動画を聴覚障害者が読唇した結果，WER は 47.7%となった．また，トレーニングデータに使用した話者と同じ話者(Overlapped Speakers)の動画を使ってテストした結果は，CER が 1.9%，WER が 4.8%となった．どちらも高い精度での自動読唇が実現されている．

3.3 LipNet の構造

3.3.1 LipNet 構造の概要

LipNet は文章レベルの学習を前提にしているため，時系列データを扱うための構造をもっている．図 3.1 に LipNet の構造を示す．

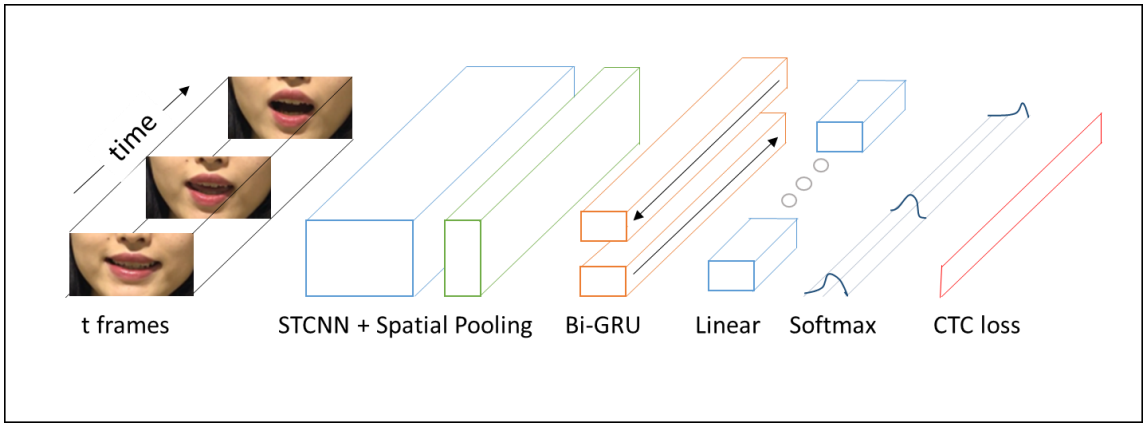

図 3.1 LipNet 構造[10]

入力には，動画データをフレーム分割したものが使われる．入力データは，まず Spatiotemporal Convolutional Neural Networks (STCNN) の三つのレイヤによって処理される．STCNN によって抽出された特徴は，二つの Gated Recurrent Unit (GRU) で処理される．LipNet では双方向に処理を行う Bi-GRU が使われており，Bi-GRU は STCNN 出力を効率的に集約させるために不可欠である．GRU 出力には各時間ステップで線形変換が適用され，Connectionist Temporal Classification (CTC) ブランクで増強された語彙に対する Softmax，CTC loss が続く．すべての層で，整流線形単位 (ReLU) 活性化機能を使用している[11].

3.3.2 Spatiotemporal Convolutional Neural Networks (STCNN)

Convolutional Neural Network (CNN) は，画像上で空間的に動作する積み重ね畳み込みを含む．そのため CNN は，入力として画像を受け取る，オブジェクト認識のようなコンピュータビジョンタスクにおいて性能を向上させるのに有効である[13]．CNN は畳み込み層とプーリング層から構成されており，三つのレイヤにはそれぞれ空間的 max-pooling 層が続く．max-pooling は，小領域に対して最大値を選択するというものである．*C* チャネルから *C'*チャネルまでの基本的な二次元畳み込みレイヤは式(3.3.2.1)で表される[11].

$$[\boldsymbol{conv}(\boldsymbol{x},\boldsymbol{w})]_{c'ij} = \sum_{c=1}^{C}\sum_{i'=1}^{k_w}\sum_{j'=1}^{k_h} \boldsymbol{w}_{c'ci'j'}\boldsymbol{x}_{c,i+i',j+j'} \qquad (3.3.2.1)$$

式(3.3.1.1)において，入力 x と重みw ∈ $\mathbb{R}^{C'\times C\times k_w\times k_h}$に対して，範囲外の i, j に対して$x_{cij}$=0 を定義する．STCNN は空間的な次元とともに時間的にも畳み込むことでビデオデータを処理することができる[14][15]．したがって，CNN と同様に，式(3.3.2.2)で表される[11].

$$[\boldsymbol{stconv}(\boldsymbol{x},\boldsymbol{w})]_{c'tij} = \sum_{c=1}^{C}\sum_{t'=1}^{k_t}\sum_{i'=1}^{k_w}\sum_{j'=1}^{k_h} \boldsymbol{w}_{c'ct'i'j'}\boldsymbol{x}_{c,t+t',c,i+i',j+j'} \qquad (3.3.2.2)$$

3.3.3 Gated Recurrent Unit (GRU)

GRU は，Recurrent Neural Network (RNN) の一種であり，より多くのタイムステップに渡って情報を伝搬させる[16]．情報の制御方法を学習するためのセルとゲートが RNN に追加されており，Long Short Term Memory (LSTM) [17]に類似した構造になっている．標準の公式は，式(3.3.3.1)である[11].

$$
\begin{aligned}
[\boldsymbol{u_t}, \boldsymbol{r_t}]^{\boldsymbol{T}} &= \mathbf{sigm}\left(\boldsymbol{W_z z_t} + \boldsymbol{W_h h_{t-1}} + \boldsymbol{b_g}\right) \\
\widetilde{\boldsymbol{h}}_{\boldsymbol{t}} &= \boldsymbol{tanh}(\boldsymbol{U_z z_t} + \boldsymbol{U_h}(\boldsymbol{r_t} \odot \boldsymbol{h_{t-1}}) + \boldsymbol{b_h}) \\
\boldsymbol{h_t} &= (\mathbf{1} - \boldsymbol{u_t}) \odot \boldsymbol{h_{t-1}} + \boldsymbol{u_t} \odot \widetilde{\boldsymbol{h}}_{\boldsymbol{t}}
\end{aligned}
\qquad (3.3.3.1)
$$

$\mathrm{z} := \{z_1, \ldots, z_T\}$ は入力シーケンスであり，LipNet では STCNN の出力となる．⊙は要素ごとの乗算を表し，$\mathrm{sigm(r)} = 1/(1 + \exp(-r))$である．LipNet の構造には，双方向に GRU の処理を行う Bi-GRU が用いられている．Bi-GRU を用いることで，単語レベルではなく，文章レベルの発話を認識することに対して頑健なネットワークの構成を実現している．

3.3.4 Connectionist Temporal Classification (CTC)

CTC loss [17] は，入力をターゲット出力に合わせるための訓練データが不必要なため，音声認識において広く使用されている．CTC では，特別な”blank”トークンで補強されたトークンクラス (語彙) にわたって一連の離散分布を出力するモデルを考える．CTC は，このシーケンスと等価であると定義されるシーケンスすべてに渡ってマージナル化することにより，シーケンスの確率を計算する．シーケンスの確率を計算することにより，アライメントの必要性を排除し，可変長配列に対処することができる．

3.4 むすび

英語話者の自動読唇システムである LipNet は文章単位の学習をすることで，高い認識精度を実現している．本章では，LipNet について詳しく述べた．

第 4 章 データ作成

4.1 まえがき

本章では，作成した日本語のデータセットについて詳しく述べる．

4.2 日本語の特性

LipNet では，英語の発話に対して高い認識精度を実現している．英語には母音が 26 個あるため，母音が五つの日本語に比べ，口の形の違いによって見分けることが容易である．英語の母音の一覧を表 4.1 に示す．

表 4.1 英語母音一覧[19]

発音記号	母音区別	単語の例
æ	短母音	apple
ʌ	短母音	fun
a	短母音	box
a:	長母音	father
aːr	長母音	arm
əːr	長母音	learn
ə	短母音	about
ər	短母音	doctor
i	短母音	it
iː	長母音	easy
u	短母音	book
uː	長母音	noon
e	短母音	egg
ɔː	長母音	ball
ɔːr	長母音	more
ai	二重母音	nice
aiər	三重母音	fire
au	二重母音	now
auər	三重母音	our
ei	二重母音	make
ɔi	二重母音	boy
ou	二重母音	open
juː	長母音	new
iər	二重母音	near
uər	二重母音	sure
eər	二重母音	air

英語の母音は，短母音，長母音，二重母音，三重母音に分類される．短母音は短く発音されるもので，長母音は，長音記号(:)が含まれている[19]．二重母音は二つの短母音から構成されており，三重母音は三つの短母音から構成されている．日本語の母音は，/a/，/i/，/u/，/e/，/o/の五つのみである．また，日本語は子音の数も少ない．日本語 45 音の一覧とその口形を図 4.1 に示す．

図 4.1 日本語 50 音の口形

図 4.1 から分かるように，日本語では母音が少ない上に，子音による大きな口形の違いも見られない．また英語には，日本人が同じ音だと感じる音でも，英語を話す人々には異なる音に聞こえる子音も存在する．さらに英語では，子音で終わる単語も存在するが，日本語では母音で終わる単語しか存在しない．そのうえ，英語では同じ音の単語を発話する際，違いを出すには音の強弱，つまりアクセントを使う．一方で，日本語では音の高さ，つまりピッチで違いを出す. 以上の理由により，日本語は単純な口形変化では見分けがつきにくい[20]. 重ねて，英語では文章を発話した際に，隣り合った単語をつなげて発話することがある．このことは，英語において文章単位で学習することが有効である理由の一つであると考えられる．日本語では，同じ音でも漢字の表記などによって違う意味を表すものが多い．文法などを考慮して文章レベルでの学習をすることでよって，意味の合っている文章が出力される可能性がある．したがって，日本語でも文章レベルの学習を行うことは重要であると考えられる．

4.3 日本語のデータベース

日本語の自動読唇のためのデータセットは少なく，公開されている単語レベルのデータセットでは, 齊藤らの Speech Scene database by Smart Device (SSSD) [21][22][23] が存在する．これは，表 4.2 に示す日本語の単語 25 語をスマートデバイスによって話者自身で撮影したものである．データは，発話動画をフレーム分割し 300x300 ピクセルのサイズに口唇周辺を抽出した画像データである．動画のフレームレートは 30fps であり，1 発話分に対するフレーム数にはばらつきがある．図 4.2 に，SSSD の画像データの例を示す．

表 4.2 SSSD の発話内容[21][22][23]

#	発話内容	#	発話内容	#	発話内容
0	ぜろ	10	ありがとう	20	どういたしまして
1	いち	11	いいえ	21	はい
2	に	12	おはよう	22	はじめまして
3	さん	13	おめでとう	23	またね
4	よん	14	おやすみ	24	もしもし
5	ご	15	ごめんなさい		
6	ろく	16	こんにちは		
7	なな	17	こんばんは		
8	はち	18	さようなら		
9	きゅう	19	すみません		

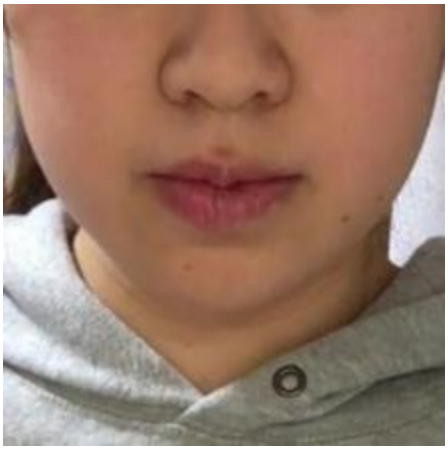

図 4.2 SSSD の例[21][22][23]

予備実験として，SSSD を LipNet に適用する．SSSD に含まれる，「0」から「5」の数字の発話シーン 992 をトレーニングデータとし，SSSD の画像データからさらに唇を中心に 100x50 ピクセルのサイズに切り抜いた．学習は 83epoch で終了した．SSSD に含まれていない「0」から「5」の発話シーンを新たに 50 シーン撮影し，テストデータとした．この結果，認識率は表 4.3 のようになった．

表 4.3 数字発話シーンテスト結果

input \ output	0	1	2	3	4	5	認識率 [%]
0	5			1	1	1	62.5
1	2			2	1	3	0.0
2	1	3		2	1	1	0.0
3	3	2		1	1	1	12.5
4	1	1		2	2	2	25.0
5	1	2		2	1	2	25.0

表 4.3 より，認識率は非常に低く，0.0%になった数字もある．LipNet は文章レベルの自動読唇を前提として作られているため，このような結果になったと考えられる．そのため，日本語の文章レベルのデータセットが必要である．

4.4 データ作成

4.2, 4.3 より，日本語の文章レベルのデータセットが必要であると考えられるが，日本語の文章レベルのデータセットは存在しない．よって，データセットの作成をする必要がある．本研究では，2 種類のデータセットを作成した．

一つ目は，日本語の文章の発話シーンデータを 360 作成した(データセット A)．発話文は，SSSD の発話内容に含まれている単語を中心に，表 4.4 に示す 18 の単語から五つの単語を組み合わせて作成した．また，LipNet で使用された GRID コーパスで採用されている文法を参考にして語順を決めた．発話シーンはスマートフォンで撮影し，動画のフレームレートは 30fps である．Dlib の顔検出器と iBug の顔ランドマーク予測子によって，発話映像から口唇周辺を 100x50 ピクセルのサイズに抽出し，フレーム分割する．すべての発話映像は 115 フレームで構成される．さらに発話映像に対応した，単語の発話区間を示すアライメントを作成する．作成したデータセットの一部を図 4.3 に示す．また，作成したアライメントの一部を図 4.4 に示す．アライメントには，それぞれの単語が発話されているフレームの区間が示されている．

表 4.4 データセット A に使用した単語

語順	1	2	3	4	5
	おはよう	あか	ぜろ	はい	ありがとう
	おやすみ	あお	いち	いいえ	どういたしまして
		きいろ	に		ごめんなさい
			さん		もしもし
			よん		おめでとう
					すみません

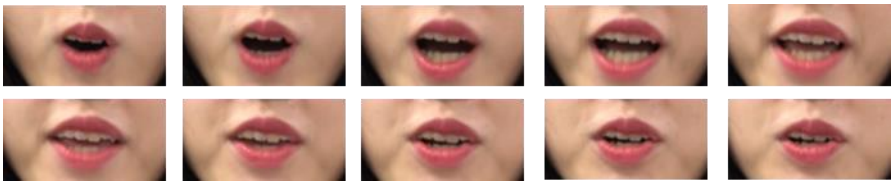

図 4.3 データセット A の一部 「おはよう」の発話シーン

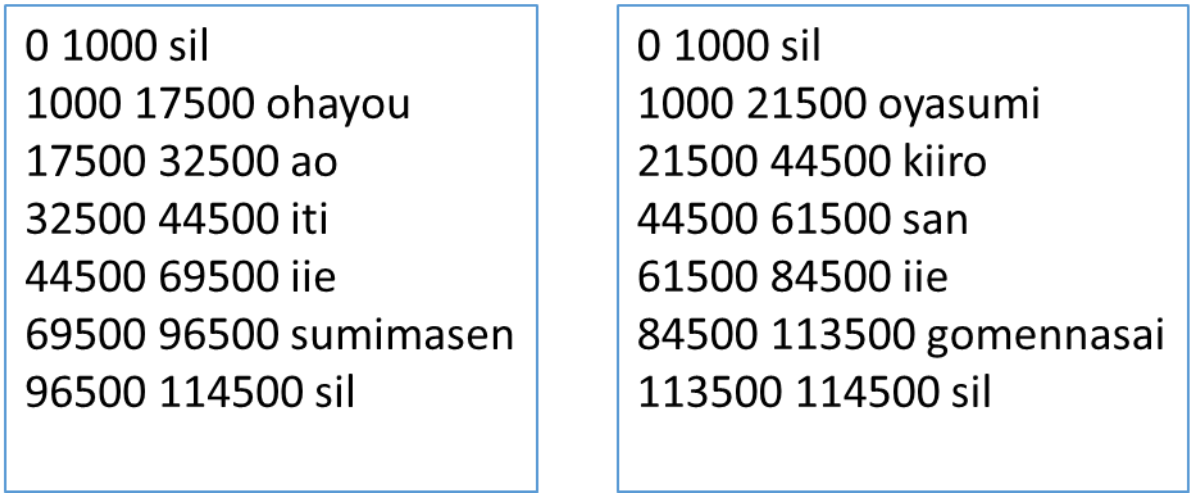

図 4.4 データセット A アライメントの一部

英語は，「p」，「b」，「m」などの，一度口を閉じて発話するという動きをもつ文字を多く含んでいる．「p」，「b」，「m」を含んだ単語は，GRID コーパスにも多く含まれている．日本語で「p」，「b」，「m」を子音にもつ音素の口形変化を図 4.5 に示す．

図 4.5 「p」,「b」,「m」を子音にもつ音素の口形変化

図 4.5 より，わずかではあるが，それぞれの音素発音時の口形の時系列変化に違いが見られる．よって子音が「p」,「b」,「m」である単語を多く含むデータセットならば認識精度が上がると考えられる．そのため,「p」,「b」,「m」を含む単語によって構成された文の発話シーンデータを 360 作成した(データセット B)．発話した文に使用した単語を表 4.5 に示す．データセット B もデータセット A と同様に，発話映像に対応した単語の発話区間を示すアライメントを作成する．

表 4.5 データセット B に使用した単語

語順	1	2	3	4	5
	あそびば	ばしょ	めざまし	またね	ぱりぱり
	ぜんぶ	びんぼう	もしもし	みんな	ぴったり
		ぶきみ	ぼくめつ	むこう	ぷっくり
		べんとう			ぺたぺた
					ぽきぽき

4.5 むすび

日本語の発音は英語とは大きく異なるため，日本語独自のデータセットが必要である．また，LipNet は文章レベルの学習をするために作られているため，本研究では日本語の文章レベルのデータセットを作成した．本章では，作成した日本語のデータセットについて述べた．

第 5 章 実験と結果，考察

5.1 まえがき

本章では，作成した日本語のデータセットを LipNet 構造に対して適用し，その結果と考察を述べる．

5.2 実験

第 4 章において，日本語の文章レベルのデータセットを二つ作成した．まず，データセット A をトレーニングデータとし，LipNet 構造に適用する．トレーニングは，損失関数の収束を基に，150epoch 繰り返した．テストデータには，表 4.4 を基に作成した文章の発話シーンを新たに 70 シーン撮影したものを使用した．テスト動画は，スマートフォンで撮影した動画をそのまま使う. データセット B も同様に, トレーニングデータとして使用し, 150epoch でトレーニングを終了した．テストデータには，表 4.5 の単語を基に作成した文章の発話シーンを新たに 70 シーン撮影したものを使った．データセット A に対するテストデータと同様に，スマートフォンで撮影したものを使った．また，データセット A，データセット B それぞれに対して作成したテストデータ（テストデータ A，テストデータ B）の読唇を，被験者に実行させた．被験者による読唇は 2 パターン行い，パターン 1 では，発話内容を全く知らせず，読唇させた．パターン 1 は 4 人に読唇させ，パターン 2 は，表 4.4.1，表 4.4.2 それぞれを見せ，表をもとに 1 人に発話内容を推測させた．

5.3 結果，考察

5.2 において行った実験の結果を述べる．図 5.1 に，データセット A に対する認識結果の一部を示す．図 5.2 に，データセット B に対する認識結果の一部を示す．また，表 5.1 に，それぞれのデータセットを使った際のテスト結果を示す．ここでは，Word Error Rate (WER), Character Error Rate (CER)それぞれの結果を算出した．さらに表 5.2 に，テストデータを人間が読唇した結果を示す．パターン 1 は WER と CER を算出し，パターン 2 は WER のみを算出した．

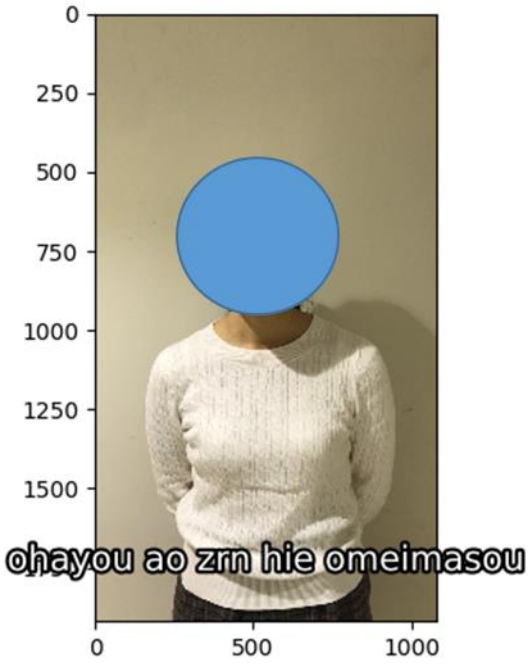

図 5.1 データセット A 認識結果の例

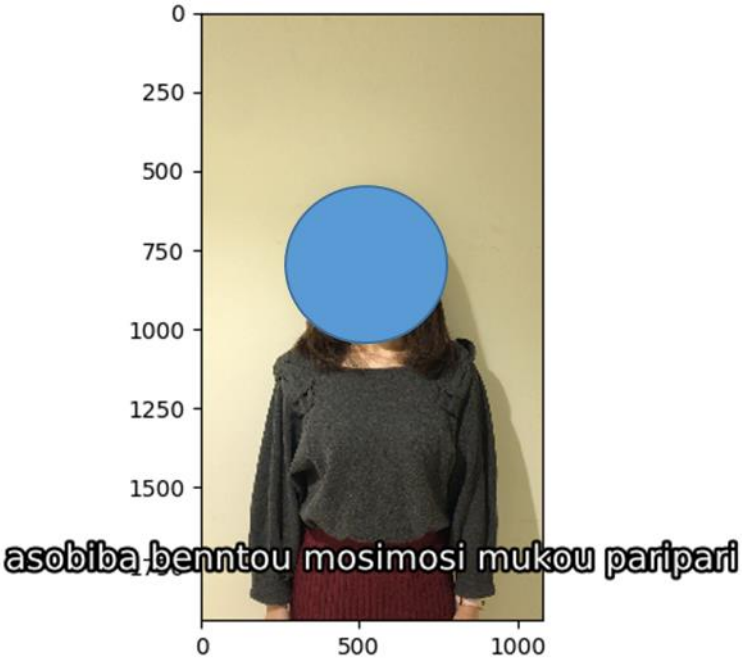

図 5.2 データセット B 認識結果の例

表 5.1 テスト結果

データセット	WER [%]	CER [%]
データセットA	81.18	68.73
データセットB	65.07	60.67

表 5.2 被験者による読唇結果

	テストデータA		テストデータB	
	WER [%]	CER [%]	WER [%]	CER [%]
パターン1-1	96.76	93.09	100	97.15
パターン1-2	90.29	84.68	96.62	94.53
パターン1-3	96.18	93.53	100	98.65
パターン1-4	85.92	82.47	100	98.01
パターン2	45.88	-	49.01	-

データセット A を使用してトレーニングした結果，表 5.1 に示すように，WER は 81.18%となり，認識精度が高くないことがわかる．データセット B を使用すると，WER は，65.07%と，データセット A を使用した場合より低くなり，認識精度が向上した．また，CER においても，データセット B を使用した場合の方が認識精度が高い．被験者による読唇では，パターン 1 においてはテストデータ A，B 共に非常に精度が低い．一方で，パターン 2 における正解率は約 50%である．発話内容のヒントを予め与えられた場合は正解率が上がるが，事前知識なしに読唇すると非常に精度が低く，機械による自動読唇は有効であると考えられる．しかし，データセット A，B のどちらを使用した場合でも，読唇は不完全である．この理由としては，第一に，トレーニングデータが不十分であることが挙げられる．膨大なトレーニングデータを使って学習することで，読唇精度が上がると考えられる．また，今回のトレーニングデータでは日本語の音すべてを取り入れることが出来なかったため，学習に使用した音素以外の認識は難しいと考えられる．よって，トレーニングデータに日本語の全ての音をバランスよく取り入れることが必要である．さらに，日本語の特性を考慮した学習モデルの構築も必要である．

5.4 むすび

本研究では，日本語の文章レベルのデータセットを作成した．本章では，作成したデータセットを LipNet 構造に対して適用し，その結果と考察を述べた．

第 6 章 結論と今後の課題

6.1 結論

本研究では，日本語の文章レベルのデータセットを作成し LipNet に適用することで，LipNet 構造の日本語のデータセットへの有効性を調べた．作成したデータセットは，LipNet で使われている GRID コーパスと日本語の単語のデータセットである SSSD を参考に作成したデータセット A と，「p」，「b」，「m」を含む単語を用いて作成したデータセット B の 2 種類である．データセット A に LipNet 構造を適用した結果，単語の認識率は，18.82%であった．また，データセット B に LipNet 構造を適用した結果，単語の認識率は，34.93%であった．同一の文章を被験者に読唇させた場合，読唇精度は非常に低く，機械による自動読唇は有効であると考えられる．しかし，自動認識の精度は不十分でありデータ数も不足しているため，日本語の完全な読唇には至っていない．

6.2 今後の課題

本研究では，LipNet を用いた日本語話者の自動読唇の手法について検討した．日本語のデータセットを作成し，LipNet 構造に適用したが，高い認識精度を得ることはできなかった．この原因としては，トレーニングデータの数が少ないことが挙げられるため，トレーニングデータ数を増やす必要がある．また，トレーニングデータを作る際には，日本語の音素すべてを学習させるために，音素バランスを考えたデータセットの作成が必要である．さらに，日本語は唇の動きのみでの発話内容の認識が難しい言語であるため，口周辺の動きや表情も学習に導入することで，認識精度が上がる可能性があると考えられる．

謝辞

本研究にあたり，環境を整えてくださり，研究についてアドバイスをくださった渡辺教授に感謝申し上げます．また，研究の進め方についてご指導くださり，お手伝いいただいた早稲田大学国際情報通信センターの石川孝明様に心から感謝いたします．さらに，データ集めなどに協力していただいた研究室の皆様にお礼申し上げます．

また，学業に専念するための環境をくださり，生活を支えてくださった家族に感謝申し上げます．

参考文献

[1] 駒井，宮本，滝口，有木，”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)，pp.1771-1778, July 2010
[2] 柿原，滝口，有木，三谷，大森，中園，“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, pp．197-200, Mar. 2015
[3] 宮崎，中島，”日本語発話時の口形変化量の分析と発話映像自動生成への適用”，マルチメディア，分散，協調とモバイル(DICOMO2013)シンポジウム，July 2013
[4] 宮崎，中島，“口形ベースの機械読唇における単語認識手法の提案と評価”，マルチメディア，分散，協調とモバイル(DICOMO2014)シンポジウム，July 2014
[5] 宮崎剛，”機械読唇による発話障碍者向けコミュニケーション支援アプリの開発”，http://www.airpf.or.jp/kenkyu/27/j07.pdf.
[6] 宮崎，中島，”日本語発話時の特徴的口形のコード化と口形変化情報表示方法の提案”，電気学会論文誌 C，Vol.129，No.12，pp.2108-2114, Dec. 2009
[7] 高橋，大谷，”複数画像特徴量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，Vol.2914-CVIM-191 No.7, Mar. 2014
[8] 義平，有馬，奥野，”フーリエ記述子法による「読唇技術」開発の試み –唇の輪郭形状による音の推定-”，人間工学 第 42 巻 特別号，June 2006
[9] パリアスカケンジ，”Deep Learning による読唇システム”，中京大学 情報理工学部 機械情報工学科 2015 年度卒業論文，Jan. 2016
[10] パリアスカケンジ，”深層学習を用いた読唇システム”，情報処理学会第 79 回全国大会，pp.557-558, Mar. 2017
[11] Y. Assel, B. Shillingford, S. Whitesaon, and N. de Freitas, “LiPNeT: End-to-End Sentence-Level Lipreading”, https://arxiv.org/abs/1611.01599, Dec. 2016.
[12] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: The first facial landmark localization challenge”, In IEEE International Conference on Computer Vision Workshops, pp.397–403, Dec. 2013.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, In Advances in neural information processing systems, pp.1097–1105, Dec. 2012.
[14] Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks”, In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp.1725–1732, June 2014.
[15] S. Ji, W. Xu, M. Yang, and K. Yu, “3d convolutional neural networks for human action recognition”, IEEE transactions on pattern analysis and machine intelligence, Vol.35, No.1, pp.221–231, Mar. 2013.
[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural

networks on sequence modeling", arXiv preprint arXiv:1412.3555, 2014.
[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural computation, Vol.9, No.8, pp.1735–1780, Aug. 1997.
[18] A. Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks", In International Conference on Machine Learning, pp. 369–376, June 2006.
[19] 英語学習ブログ，"英語の発音の学習。母音と子音は、いくつあるの？", https://english-learning-blog.hatenablog.com/entry/boin-siin. 2015
[20] 英語びより，"日本語と英語が言語的に違いすぎる!だから習得が難しい", https://ipa-mania.com/japanese-english-difference/. 2018
[21] 齊藤, 窪川: "SSSD：スマートデバイスを用いた読唇技術向け日本語データベース"，信学技報, PRMU2017-199, Mar. 2018
[22] 窪川，齊藤: "SSSD を用いた深層学習による読唇精度に関する検討"，第 21 回 画像の認識・理解シンポジウム (MIRU2018) ,PS2-39, Aug. 2018
[23] T. Saitoh, and M. Kubokawa: "SSSD: Speech Scene Database by Smart Device for Visual Speech Recognition", Proc. of International Conference on Pattern Recognition, pp.3228-3232, Aug. 2018.

図一覧

表一覧

研究業績

[1] 浅見，石川，渡辺，”機械学習による日本語話者の自動読唇の基礎検討”，第 33 回画像符号化シンポジウム・第 23 回映像メディア処理シンポジウム（PCSJ/IMPS2018），P-3-08，Nov. 2018.

[2] 浅見，石川，渡辺，”機械学習による日本語話者の自動読唇”，電子情報通信学会総合大会，Mar. 2018(発表予定).

[3] 横井，浅見，石川，渡辺，”ボロノイ図に基づく 3 次元優勢領域によるパスコース評価につ