機械学習による日本語話者の自動読唇の基礎検討

Basic Study on Lip Reading for Japanese Speaker by Machine Learning

浅見 莉絵子† 石川 孝明†‡ 渡辺 裕†‡
Rieko Asami† Takaaki Ishikawa†‡ Hiroshi Watanabe†‡

早稲田大学基幹理工学部情報通信学科†
School of Fundamental Science and Engineering, Waseda University†

早稲田大学国際情報通信センター‡
Global Information and Telecommunication Institute, Waseda University‡

1． まえがき

近年，機械学習の応用分野として自動読唇の研究が行われている．こ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，自動読唇の研究は発展途上にある．本研究では，英語の文章レベルでの自動読唇において注目されている LipNet を用いた日本語話者の自動読唇の手法について検討する．

2． 日本語読唇の従来手法

日本語の自動読唇手法には，Active Appearance Model (AAM)を用いた手法がある．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．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 Nueral 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，どちらの手法も検出精度が低く，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

我々は，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ことを試みる．

LipNet は、英語話者の自動読唇システムであり，英語発話に対する検出精度は 93.4%である．図 1 に LipNet の構成を示す．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]．これ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．このモデルは，Connectionist Temporal Classification (CTC)で訓練されている．

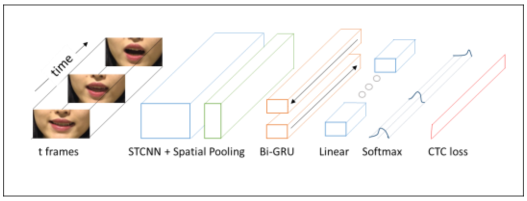

図 1 LipNet architecture [3]

4． 実験

LipNet で日本語のデータセットを学習させ，日本語話者に適用する．本実験では，表 1 に示す 20 の単語から 5 つずつ組み合わせた文の発話シーン 360 を学習させ，未知データ 36 を評価した．今回は，エポック数 105 回で学習を終了した．この時の単語の認識率は 36.0%であった．この原因として，学習データが不十分であることが考えられる．また，日本人は発話時に口を大きく開かないため，口の形の変化があまり表れなかったことが考えられる．

表 1　学習に使用した単語

語順	1	2	3	4	5
	おはよう	あか	ぜろ	はい	ありがとう
		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
		みどり	さん		もしもし
		しろ	よん		おめでとう
			ご		すみません

4． むすび

本稿では，LipNet の日本語への適用を行った．今回の実験では，学習データが不十分であり，従来法を凌ぐ結果は得られなかった．今後の検討事項として，口周辺の動きや表情の学習への導入などが考えられる．

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, 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, 2015.

[3] Y. Li, Y. Takashima, T. Ttakiguchi, and Y. Ariki, ”Lip Reading Using a Dynamic Feature of Lip Images and Convolutional Neural Networks”, IEEE ICIS, pp.1-6, June 2016.

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