#077

JasperNet：顔動画像解析による日本語の発話予測

東京都立大泉高等学校　大泉データサイエンス樫　佐々木俊輔

要旨：

Lipreading in this study refers to the activity of understanding speech content from the movements of the mouth and its surroundings. The goal of this study is to improve the accuracy of Japanese lipreading by deep learning. Compared to English lipreading, Japanese one is more difficult, so it is of great social and academic significance to demonstrate its effectiveness. The training data was devised because Japanese has many throat sounds. A neural network structure was devised for delicate feature extraction of face images and efficient learning of mouth shape transitions. We concluded that this specially designed JasperNet showed higher accuracy than conventional methods, and that these innovations contribute to the improvement of lipreading accuracy.

背景：

音声認識の課題：「雑音下」「公共の場」「障がい者」

発話認識：(駒井ら, 2010)
音声による精度 72.4　%
画像による精度 11.85 %

AI読唇（LipNet）：
英語　95.2 % (Assaelら, 2016)
日本語 44.1 % (北原ら, 2021)

目的・意義：

高精度で日本語の読唇を行う深層学習モデルの開発

学術的意義：
他言語への応用
読唇の研究の活性化

社会的意義：
障がい者支援，経済効果
諜報機関，公共の場

材料・方法：

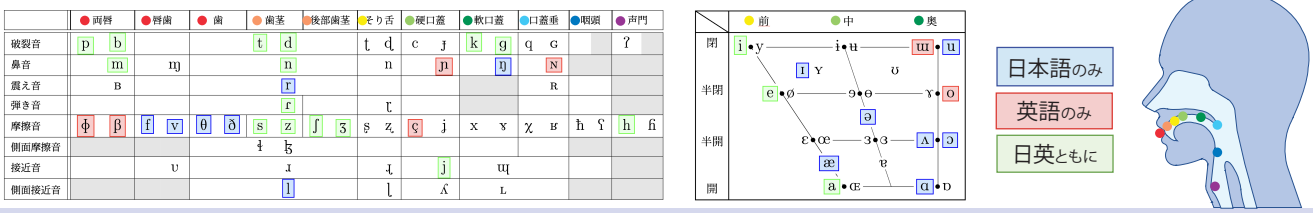
日本語発音の分析：
・IPA（図表1）
・音節拍リズム

データの準備：
・撮影(HD, 90フレーム, 10000件)
・切り取り（DLib・iBug)

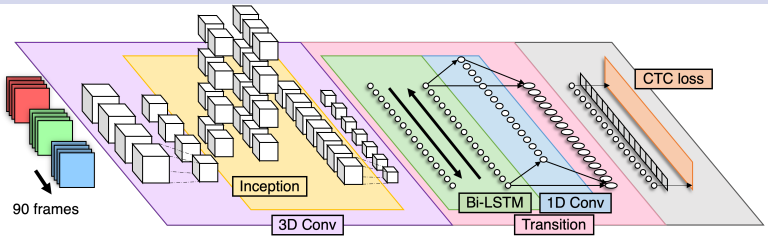
ニューラルネットワーク：
・読唇モデルの開発
・日本語の特徴の考慮

実証評価：
・顕著性マップ
・CTC loss比較

図表 1 日本語と英語の音素（左:子音 右:母音）



結果・考察：



CNN多層化：
繊細な特徴抽出

表現力向上：
LSTMでパラメータ増

勾配消失対策：
並列に処理

口形遷移に注目：
時間方向に畳み込み

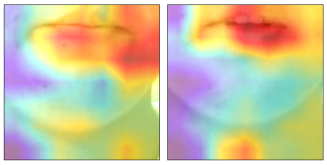


図 2 Grad-camによる顕著性マップ

表 2 CTC損失値の比較

方法	損失
唇のみ＋LipNet 構造	3.13
唇・喉＋LipNet 構造	2.97
唇のみ＋JasperNet 構造	2.31
唇・喉＋JasperNet 構造＝JasperNet	2.10

結論・展望：

・JasperNetによる読唇の実現可能性
・高い汎化性能の獲得・モデル公開
　→**新しいコミュニケーション**

引用文献：

[1] 駒井拓人, 宮本千琴, 滝口哲也, & 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. 画像の認識・理解シンポジウム, MIRU2010, IS3-31, 1771-1778.
[2] Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599.
[3] 北原瑠伊, & 猪力峰. (2021). 機械学習を用いた日本語読唇における五十音データセット作成の提案. In 産業応用工学会全国大会講演論文集 2021 (pp. 28-29). 一般社団法人 産業応用工学会.
[4] 泉谷双葉. (2012). もうひとつの五十音図. 東京医科歯科大学教養部研究紀要, 42, PAGE47-62.
[5] IPA Chart(2022). IPA Chart. https://www.ipachart.com. 2022 年 10 月 23 日
[6] D. Easton and M. Basala. Perceptual dominance during lipreading. Perception & Psychophysics, 32(6): 562–570, 1982.
[7] Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. The Journal of the Acoustical Society of America, 123(4), 2356-2366.
[8] King, D. E. (2009). Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10, 1755-1758.
[9] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In Proceedings of the IEEE international conference on computer vision workshops (pp. 397-403).
[10] Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.
[11] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1- 9).
[12] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[13] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11), 2673-2681.
[14] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
[15] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
[16] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376).
[17] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning (pp. 173-182). PMLR.

謝辞：

本研究の遂行にあたり，指導教員として終始多大なご指導を賜りました，山口貴史先生に深謝致します。廣瀬拓海氏，並びに猪野凱氏には，ティーチングアシスタントとして適切なご助言を賜りました。ここに深謝の意を表します。最後に，友人 ・ 家族には，長時間にわたる撮影にご協力頂きました。ここに感謝の意を表します。