調理音とレシピテキストを用いた調理工程の推定

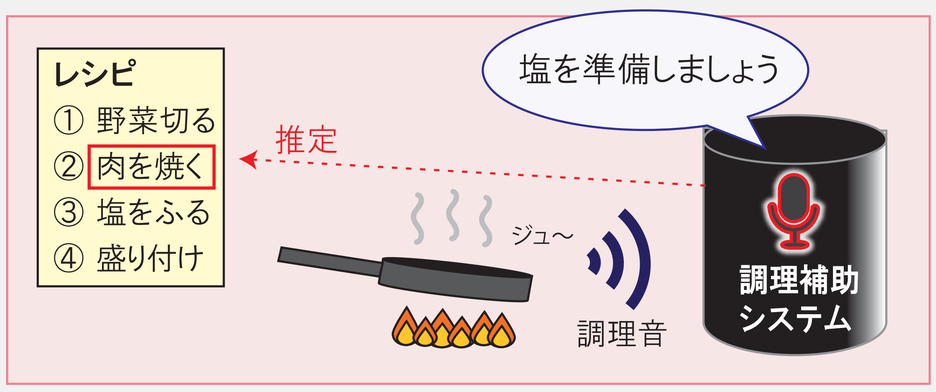
岩川光一　井本桂右（同志社大学）　S06

1. 研究動機

調理工程の推定を行う理由

・料理初心者がレシピ通りに調理する時

・調理工程を **推定** できれば…

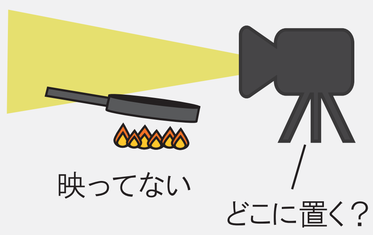


2. 従来研究

音以外を用いた調理作業の推定

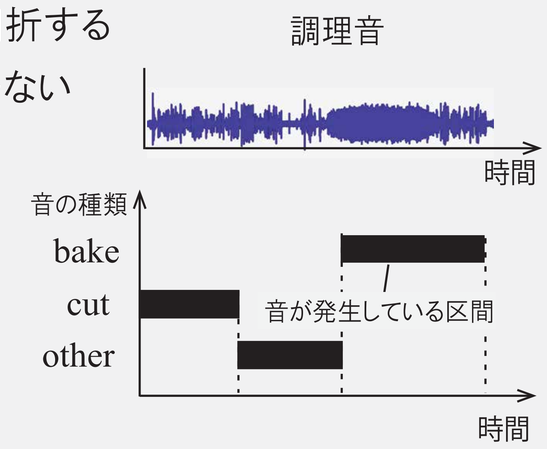
・画像や映像を用いて食材の調理状態を推定 [1]
・モーションキャプチャやジャイロセンサを用いて料理の種類と作業の種類を推定 [2]

短所：　センサの設置や装着が面倒



音を用いた調理作業の推定

音の特性：　指向性が広い、回折する
→センサ設置位置の制約が少ない



先行研究

・調理音の中で、「焼く」、「切る」、「その他」それぞれの音が鳴っている区間を推定 [3]

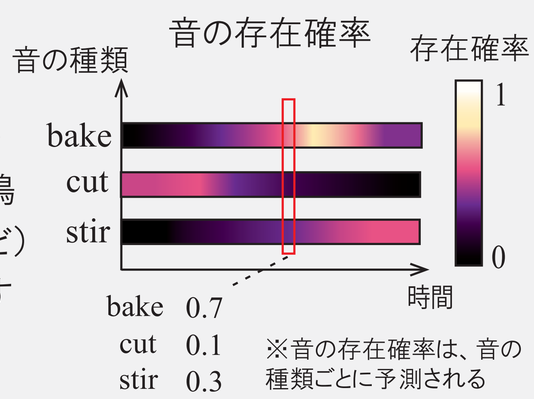
先行研究の課題

① 「焼く」、「切る」といった調理作業そのものだけを推定できたとしても、調理補助に使いづらい

② 推定できる作業の種類が少ない

環境音の存在確率の予測

環境音の事前学習済みモデル [4] を用いて、身の回りに存在する音（鳥の鳴き声、水音、エンジン音、食器の音など）が各時点で存在している確率を予測することが可能



3. 提案手法

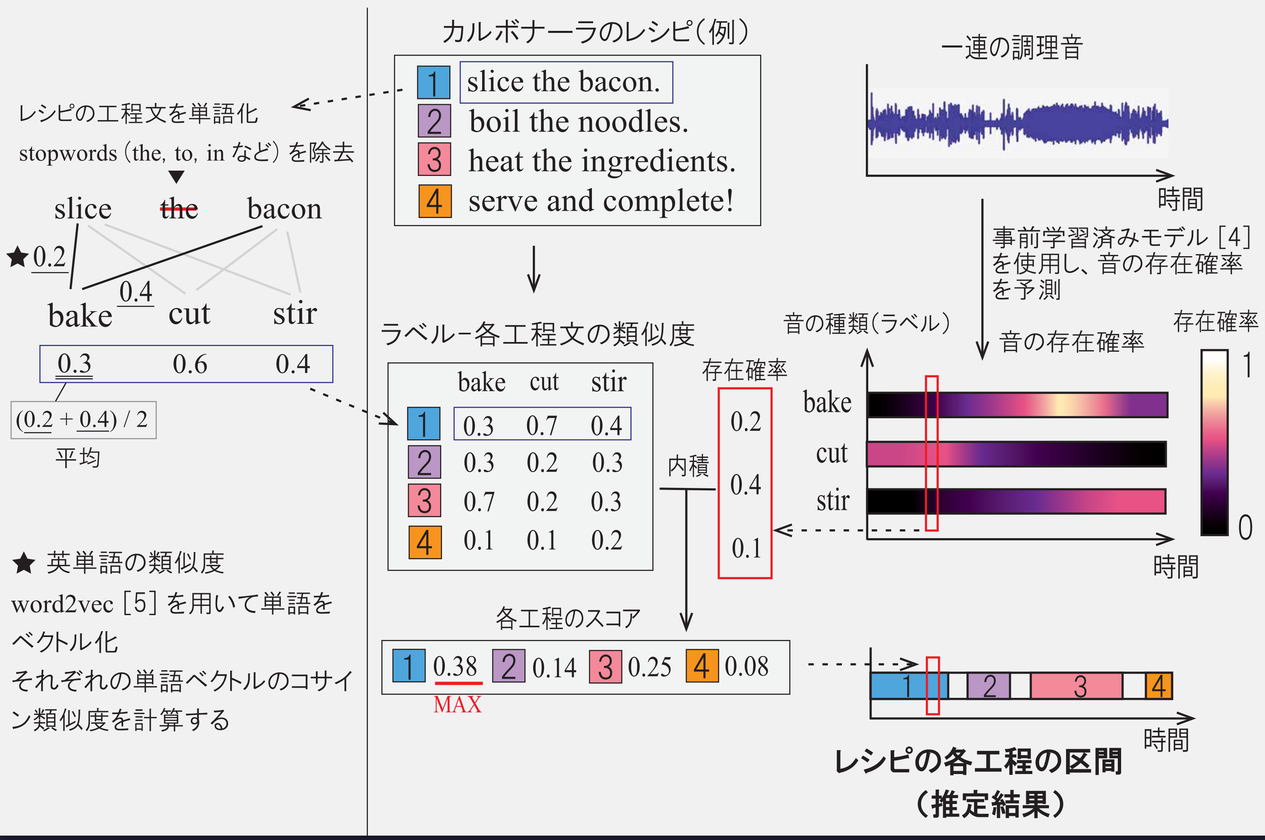
本手法のポイント

① レシピの工程を推定
→　次の工程の内容を示すなど、調理補助に活用

② 環境音の事前学習済みモデル [4] を使用
→　食器の音や電子レンジの音など、日常生活で発生する様々な音を検出し、多種の調理工程を推定

実装の説明

・調理の始まりから終わりまでの一連の音に対し、レシピの各工程の区間を推定



★ 英単語の類似度
word2vec [5] を用いて単語をベクトル化
それぞれの単語ベクトルのコサイン類似度を計算する

4. 評価実験

前提条件

− 使用した音
・YouTube のカルボナーラの調理動画 [6] の音
・工程と工程の間にカット有

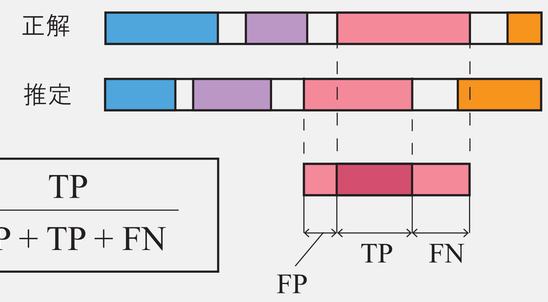
使用レシピ（カルボナーラ）

1. Cut the bacon.
2. Grate the cheese and mix it with eggs.
3. Boil the noodles
4. Fry the bacon.
5. Add the boiled noodles and mix.
6. Plate it , and it's ready to serve

− 使用したレシピ
・上記の調理動画に対して自分で書き起こし、chatGPT で英訳したもの

評価指標

・mean Intersection over Union (mIoU)
各工程の IoU の平均



$$IoU = \frac{TP}{FP + TP + FN}$$

実験結果

工程	1	2	3	4	5	6
IoU	0.00	0.00	0.70	0.00	0.42	0.00

mIoU:　0.19

FP (False Positive)	工程の区間ではない部分を誤って工程の区間であると推定した
TP (True Positive)	工程の区間である部分を正しく推定できた
FN (False Negative)	工程の区間である部分を推定できなかった

参考文献

[1] Salekin, Md Sirajus, Ahmad Babaeian Jelodar, and Rafsanjany Kushol. "Cooking state recognition from images using inception architecture." 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 2019.
[2] Cooking Activity Recognition Challenge (https://abc-research.github.io/cook2020/)
[3] Yusaku Kerematsu, Daisuke Saito, and Nobuaki Minematsu. "Cooking State Recognition based on Acoustic Event Detection." Proceedings of the 11th Workshop on Multimedia for Cooking and Eating Activities (CEA) (2019): 41-44
[4] Kong, Qiuqiang, et al. "Panns: Large-scale pretrained audio neural networks for audio pattern recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 2880-2894.
[5] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
[6] Emojoie えもじょわ 「本格！カルボナーラの作り方｜えもじょわキュイジーヌ」(https://www.youtube.com/watch?v=3Tv935A-vEw&t=1s&ab_channel=Emojoie)