6N-07

マルチチャンネルの音情報を用いた咀嚼と嚥下の自動検出

中村　亮裕†　　齊藤　隆仁‡　　池田　大造‡　　太田　賢‡　　峰野　博史‡　　西村　雅史‡

静岡大学大学院総合科学技術研究科†　　株式会社 NTT ドコモ‡

1. はじめに

咀嚼と嚥下を含む一連の摂食行動は，健康を維持する上で大変重要だと考えられている．ただ，その摂食行動の大半は人間の体内で行われる動作であり，正確な行動データを収集するには X 線やファイバースコープといった侵襲性の高い手段を用いる必要がある．保健・医療の分野でも，より簡便な測定方法が求められていることから，我々は食事の際に生じる音から一連の摂食行動を自動認識する手法の開発に取り組んでいる．これまでに Hybrid CTC/Attention model を用いることで咀嚼や嚥下といった一連の行動検出に加え，口腔内での咀嚼位置（前・左・右）の検出までもが可能になることを示したが，その検出精度についてはまだ課題が残っていた[1]．今回，耳下装着の左右マイクの情報に加え，首元装着の上下マイクと口元の接話マイクから得られる音情報の活用を試みた．その結果，特に前咀嚼や嚥下についてさらに安定した性能を得られていることを確認したので報告する．

2. 提案手法

先行研究[1]では，耳下に装着した左右 2ch のマイクによって収録された食事音を用いて咀嚼位置（前・左・右）と嚥下の自動検出を試みている．本稿では摂食行動の客観データを高精度に自動抽出することを目的として，新たに 3ch 分の音情報の追加利用を検討する．

マイクの装着例を図 1 に示した．耳下に装着したマイク（a・b）の音情報には左右咀嚼の情報が含まれる．特徴量として相互相関と信号和の MFCC を用いることが有効であることを確認している[2]．首元に装着したマイク（c・d）の音情報には嚥下の情報が含まれる．嚥下時に食塊が咽頭から食道へと移動するため，この位置にマイクを装着することによって一連の嚥下が精度良く検出できると考えられる．特徴量として上下信号の相互相関が有効であることが報告されており[3]，今回は左右咀嚼検出と同様に相互相関の抽出を行った．口元の接話マイク（e）の音情報には，最初に食材を口に入れるときに発生する前咀嚼の情報が特に含まれる．

今回は Hybrid CTC/Attention model について，Early Fusion と Late Fusion の 2 種類の手法を検討した．図 2 に示すように，Early Fusion では，耳下，首元，口元の音情報に対して特徴抽出を行い，得られたそれぞれの特徴量をすべて連結して Hybrid CTC/Attention model に入力する．一方 Late Fusion では，音情報から得られたそれぞれの特徴量を Hybrid CTC/Attention model に入力して，最後に 3 つの確率の加重和をとる．音声と画像を Late Fusion したものが Early Fusion したものに比べて比較していくつかのノイズに頑強であることが報告されている[4]．

図 1：マイクと装着例

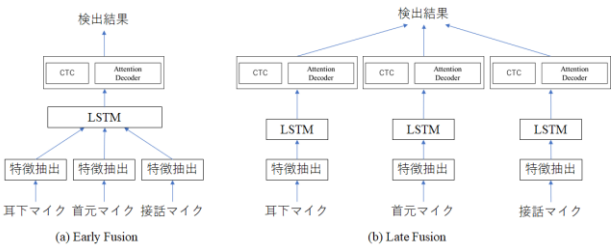

図 2:Early Fusion と Late Fusion の比較

3. 咀嚼位置と嚥下の自動検出システム

食事音を前咀嚼，左咀嚼，右咀嚼，嚥下，その他の 5 クラス分類にする．その他には無音区間やノイズが含まれる．

最初に食事音の収録（サンプリング周波数 22,050Hz，量子化 16bit）を行う．同時に学習モデルの作成のため各イベントに対し正解弱ラベルの付与を行う．弱ラベルは正確な時間情報を持たないラベルである．ラベルの付与は，オンラインアプリケーション[1]によって，被験者が食事音収録時に同時に行う．

次に収録した食事音から特徴抽出を行い，音響特徴量に変換する．特徴抽出は，フレームシフト 40ms，窓幅 80ms で抽出した．特徴量として，耳下装着の左右マイクと首元装着の上下マイクそれぞれについて，信号強調のため 2ch の信号和を取った上

Automatic Detection of Chewing and Swallowing Using Multichannel Sound Information
Akihiro Nakamura†, Takato Saito‡, Daizo Ikeda‡, Ken Ohta‡, Hiroshi Mineno†, Masafumi Nishimura†
† Shizuoka University
‡ NTT DOCOMO

で 39 次元の MFCC を求め，これを 7 次元の相互相関と連結した[2]．さらに，接話マイクの 39 次元 MFCC 特徴量も加えた．

これらの特徴量を Hybrid CTC/Attention model に入力して，咀嚼位置と嚥下の検出を行う．CTC による局所的な特徴を考慮した検出と，Attention による文脈を考慮した検出の併用により，両者の特徴を活用した認識ができる利点がある．咀嚼位置と嚥下の検出についても複雑な文脈を持つため，Attention の文脈を考慮した検出が有効であることを確認している[1]．

4. 実験

4.1. データセット

学習用データとして，クラッカー(リッツ)，キャベツ（千切り），チューインガムの食事音と，ゼリーと水 3ml，水 20ml の嚥下音を 20 代男女 30 名から収録した．同時に被験者の自己申告により弱ラベルの付与を行った．収集した音イベント数は咀嚼と嚥下で約 30,000 個となった．また評価用には学習時とは異なるオープンな食材として，ピザとりんごの食事音を 20 代男女 5 名から収録した．咀嚼と嚥下で約 1,000 個分の強ラベルの付与を弱ラベルと一部の顔画像データも参考に人手で行い，正解データとした．

4.2. 評価

全体の性能評価には，フレーム単位で推定した Mean Absolute Percentage Error（MAPE[4]）を用いた．またイベントごとの性能評価には，イベント単位の再現率，適合率，F 値を用いた．

4.3. 実験結果と考察

全体の検出性能を表 1 に示す．従来の耳下マイクの音情報のみを用いた場合に比較して，首元マイクや接話マイクの音情報を用いることで全体的な検出性能を改善できていることがわかる．また，情報統合の方法としては Late Fusion の方が有効であることが示唆された．

食材別の検出性能と経過時間の関係を図 3 に示す．食事時間を正規化して 6 分割している．ピザやりんごはチューインガムとは異なり，食事の後半にかけて検出性能が低下していることが確認できる．被験者の自己申告をベースにして，顔画像データも用いて正解咀嚼位置を付与しているが，実際にはこれらの食材は口の中に広がり全体で咀嚼するようになり，曖昧となることが影響していると考えられる．

最後にイベントごとの検出性能を表 2 に示す．咀嚼位置については，被験者の報告により実際には食事の後半では全体で咀嚼していることが分かったため，後半 3 割は除いて評価している．首元マイクを用いることで嚥下の検出性能が向上し，接話マイクを用いることで特に前咀嚼の検出性能が向上していることを確認した．非侵襲な音情報だけで，相当に高い摂食行動認識が実現できる可能性を示すことができた．

表 1：全体の検出性能
（前咀嚼・左咀嚼・右咀嚼・嚥下・その他）

Channel	MAPE (%)	
	Early Fusion	Late Fusion
2ch（耳下）	20.3	
4ch（耳下+首元）	18.9	18.3
5ch（耳下+首元+接話）	18.6	**18.0**

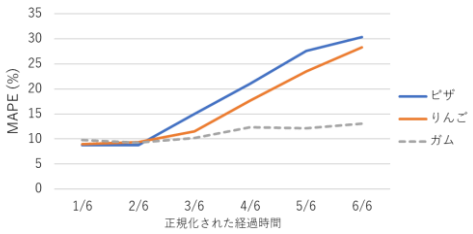

図 3： 食材別の検出性能と経過時間の関係(5ch, Late Fusion)

表 2：イベントごとの検出性能
（前咀嚼・左咀嚼・右咀嚼・嚥下・その他）

イベント	2ch（耳下）			5ch（耳下+首元+接話） Late Fusion		
	再現率	適合率	F 値	再現率	適合率	F 値
左咀嚼	0.88	0.92	0.90	0.91	0.94	0.93
右咀嚼	0.89	0.93	0.91	0.90	0.94	0.92
前咀嚼	0.46	0.64	0.55	0.61	0.80	0.71
嚥下	0.90	0.70	0.80	0.93	0.98	0.95

5. おわりに

耳下に装着した左右マイク，首元に装着した上下マイク，口元の接話マイクの音情報の併用と Late Fusion により，左右咀嚼の安定した検出性能を保ちながら，前咀嚼や嚥下の検出性能が向上することが確認でき，高精度の摂食行動認識を実現できる可能性を示すことができた．

今後は，データ量の不足に対する対応方法の検討や，食事の後半部分の評価方法の検討を行う予定である．また，Transformer や BLSTM などの別のモデルを用いる手法の検討も行う予定である．

謝辞

本研究の一部は JSPS 科研費 JP18H03260 の助成を受けたものである．

参考文献

[1] Akihiro Nakamura et al. “Automatic Detection of Chewing and Swallowing Using Hybrid CTC/Attention,” GCCE 2020, pp.333-335, 2020.
[2] Akihiro Nakamura et al. “Automatic Detection of the Chewing Side Using Two-channel Recordings under the Ear,” LifeTech 2020, pp.82-83, 2020.
[3] Yuta Sakamoto et al. “２チャンネル咽喉マイクを用いた嚥下音の認識,” 平成 30 年度電気・電子・情報関係学会東海支部連合大会, Po1-22, 2018.
[4] Stavros Petridis et al. “Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture,”, SLT 2018, pp.513-520, 2018.