# CTC と Attention の併用による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†

静岡大学情報学部† 株式会社 NTT ドコモ‡

## 1. はじめに

食事行動の質の低下は健康に悪影響を与えることが知られている．特に，一嚥下あたりの咀嚼回数が少ない人は早食いの傾向があり，肥満などの原因にもなっている．また，咀嚼位置の偏りは，歯の脱落や顔の歪みなどの原因となる．そこで我々は，健康維持の観点から重要とされている咀嚼と嚥下に着目し，一連の食事行動を簡便にモニタリングできるシステムの開発を行っている．

安藤ら[1]は咽喉マイクで収録した食事音を用いて食事行動の検出を試みたが，学習データにはフレーム単位の正確なラベル（強ラベル）の付与が必要であり，少量のデータでは十分な性能を得ることができなかった．一方，Billah ら[2]は，LSTM-CTC を使用することで，正確な時間情報の無い弱ラベルだけが付与された大量の学習データを活用し，咀嚼と嚥下の検出性能を大幅に改善できることを示した．

ここでは，先の研究を発展させ，耳下部に装着した左右 2 チャンネルの食事音を新たに利用し，嚥下や咀嚼の検出にとどまらず，左右および前方の咀嚼位置を自動検出することを試みた．なお，弱ラベルによる学習に加え，一連のイベントのコンテキスト情報を活用するため，CTC と Attention のハイブリッドモデルに基づくシステムを構築している．

## 2. 提案手法

長さが異なる入出力系列を学習させる手法として，CTC(Connectionist Temporal Classification)を用いるモデルと Attention を用いるモデルが存在する．CTC は，blank と呼ばれる空白ラベルを導入することで，長さが異なる入出力系列を学習することが可能となる損失関数である．CTC により，時系列順にそれぞれのイベントを独立に検出できる．Attention は，Encoder-Decoder モデルに導入される枠組みである．Attention では Decoder で出力する際に，Encoder の各時刻の隠れ層の状態をどの割合で利用するか，その重みを学習する．CTC とは異なり，それぞれのイベントが独立に発生することを仮定していないので，過去の任意の地点の入力を必要に応じて参照することができる．

Watanabe ら[3]は，CTC と Attention のハイブリッドモデルにより，大幅な精度改善が得られることを報告している．ハイブリッドモデルでは，CTC と Attention による出力ベクトルを足し合わせたものを最終的な出力とし，CTC と Attention 双方の特徴が生かされるメリットがある．

本稿ではこのモデルを食事行動の分析に利用する．収録した食事音を用いて，それぞれのイベントについて咀嚼位置（前・左・右），嚥下，その他（ノイズ）の 5 クラス分類を行う．Encoder には，2 層 200 次元の単方向の LSTM，Decoder には 1 層 200 次元の単方向の LSTM を用いる．CTC での各食事行動の検出に加えて，過去の複雑な食事行動の履歴を反映できる Attention の利用によってイベント検出の精度向上が期待できる．クラッカー（リッツ）を 1 枚食した時のイベント列の例を図 1 に示す．最初に前咀嚼が起こって，左右咀嚼を繰り返したあとに嚥下が行われる．咀嚼位置と嚥下を含んだ系列は，複雑なイベント系列になっていることがわかる．

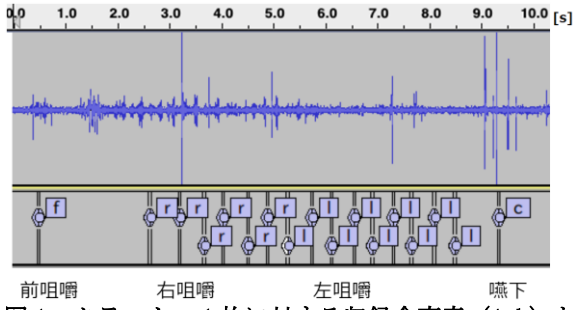


**図 1: クラッカー 1 枚に対する収録食事音（1ch）と人手で付与されたイベント列（強ラベル）の例**

## 3. 実験

### 3.1. データ収集

20 代の男女 18 名から，チューインガム，クラッカー（リッツ），キャベツ（千切り）の食事音を，耳下 2ch コンデンサマイクにより収録（サンプリング周波数 22KHz，量子化 16bit）した． 2ch マイクの装着例を図 2 に示す．2ch マイクは 3D プリンタを用いて独自で制作したものである．


**図 2:2ch マイクの装着例**

学習モデルの作成のために，収録と同時にそれぞれの咀嚼（左・右・前）と嚥下に対し弱ラベルの付与を行った．弱ラベルは正確な時間情報を持たないラベルである．ラベリングの付与のコスト削減のために，オンラインアプリケーションを作成している．被験者は咀嚼と嚥下時にキーボードのキーを押すことでイベントログの作成を行い，そのログを弱ラベルとしてモデルの学習時に使用した．

この結果，咀嚼 17,691 回（前：1,102 回，左：8,871 回，右：7,718 回），嚥下 1,982 回のデータが得られた．

---

Automatic Detection of Chewing and Swallowing Using Hybrid CTC/Attention

Akihiro Nakamura†, Takato Saito‡, Daizo Ikeda‡, Ken Ohta‡, Hiroshi Mineno†, Masafumi Nishimura†

†Shizuoka University

‡NTT DOCOMO

### 3.2. 特徴量

特徴量は，窓幅 80ms，シフト幅 40ms で抽出し，39 次元の MFCC と 7 次元の相互相関を結合したものを使用した．MFCC では，観測信号の強調処理を行うために左右の信号を足し合わせた上で，12 次元の MFCC（Mel Frequency Cepstral Coefficients）を求め，1 次元の RMS（Root Mean Square）を付加し，それらの変化量である 13 次元のΔ，13 次元のΔΔを加えた．相互相関では，前後 7 点シフトさせてそれぞれ相互相関値を求めた．また,特徴量の比較のため，左の信号の MFCC，左右の信号の和の MFCC についても評価を行う．

なお，提案手法による咀嚼位置の推定結果に対する事後処理として，1 秒未満の区間で咀嚼位置の交代が起こった場合，一方の咀嚼のみになるように修正を行い，短時間で咀嚼位置の交代が発生しないようにスムージング処理を行った．

### 3.3. 評価

性能評価は 18 名のデータに対し，話者単位で 6 分割の交差検証を行なった．評価尺度として，MAPE（Mean Absolute Percentage Error）を用いた．MAPE では，データ数を $N$，正解回数を $A_k$，推定回数を $F_k$ として，以下の式で算出する．

$$\mathrm{MAPE} = \frac{100}{N}\sum_{k}^{N}\left|\frac{A_k - F_k}{A_k}\right|$$

なお，MAPE に基づく交差検証時において収録データをテストデータとして使用する場合のみ，弱ラベルではなく，別途人手で付与した強ラベルを正解データとして使用している．

また，イベント別の検出性能としてイベント単位の再現率，適合率，F 値でも評価を行う．ここでは正解ラベルと推定ラベルの重なりが検出された場合に正しい検出が行われたとしている．

### 3.4. 実験結果と考察

咀嚼（左・右・前）と嚥下の検出性能を表 1 に示す．CTC を単独で用いたモデルと比較して，提案手法（CTC と Attention の併用）のモデルが高い検出性能を示した．また，提案手法に対するスムージングの効果は CTC 単独の方法に比べて小さく，咀嚼位置の検出においても，Attention による過去の咀嚼位置に関する履歴を考慮することが有効である．

特徴量別の咀嚼（左・右・前）と嚥下の検出性能を表 2 に示す．2ch の信号の和と相互相関を組み合わせた特徴量を用いることで，単に 2ch の信号の和を用いる場合に比べ，性能が大幅に改善されたことがわかる．

イベント別の検出性能を表 3 に示す．提案手法によって，いずれのイベントに対しても検出性能が大幅に改善された．また，左右咀嚼の検出性能は高く，前咀嚼と嚥下についてもある程度の検出ができることを確認できた．また,食材間での検出性能の違いもほとんどなく,どの食材でも安定して検出できることがわかった．

最後に Attention の重み付け結果を図 3 に示した．形状が安定しているチューインガムは，時間経過による重み付けのされ方の変化が少ない．徐々に食材が口の中に広がるクラッカーは，後半にかけて重み付けの時間の範囲が広がって，より広い行動の履歴を見るようになる傾向がある．

**表 1:咀嚼（左・右・前）と嚥下の検出性能**

| モデル | MAPE(%) |
|---|---|
| CTC | 56.2 |
| CTC（スムージング後） | 33.6 |
| **CTC+Attention** | **21.2** |
| **CTC+Attention（スムージング後）** | **19.5** |

**表 2: 特徴量別の咀嚼（左・右・前）と嚥下の検出性能 (CTC+Attention（スムージング後))**

| 特徴量 | MAPE(%) |
|---|---|
| 左信号の MFCC | 95.2 |
| 2ch 信号の和の MFCC | 33.0 |
| **2ch 信号の和の MFCC+相互相関** | **19.5** |

**表 3:イベント別の検出性能**

| | CTC（スムージング後） | | | CTC+Attention（スムージング後） | | |
|---|---|---|---|---|---|---|
| イベント | 再現率 | 適合率 | F 値 | 再現率 | 適合率 | F 値 |
| 左咀嚼 | 0.74 | 0.83 | 0.76 | 0.83 | 0.87 | 0.85 |
| 右咀嚼 | 0.76 | 0.82 | 0.78 | 0.85 | 0.89 | 0.87 |
| 前咀嚼 | 0.34 | 0.66 | 0.45 | 0.44 | 0.67 | 0.53 |
| 嚥下 | 0.80 | 0.40 | 0.60 | 0.80 | 0.70 | 0.75 |

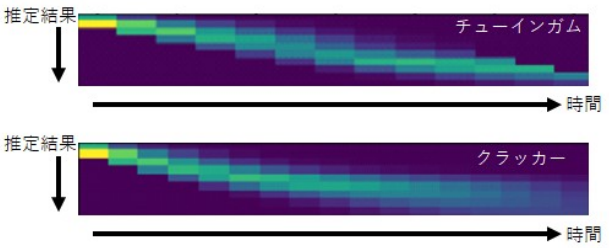


**図 3: Attention の重み付け結果**

## 4. おわりに

CTC と Attention を併用したモデルを使用することにより，咀嚼・嚥下の検出に加え，咀嚼については前・左・右の咀嚼位置の推定が可能となる見通しが得られた．

今後は実食事データでの評価による実用性の検証および食事行動の可視化についての検討を行う予定である．

## 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] Jumpei Ando et al. "Dietary and Conversational Behavior Monitoring by Using Sound Information", NCSP 2018, pp.675-678, 2018.

[2] Muhammad Mehedi Billah et al. "Estimation of Number of Chewing Strokes and Swallowing Events by Using LSTM-CTC and Throat Microphone", Proc. of GCCE 2019, pp.944-945, 2019.

[3] Shinji Watanabe et al. "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition", IEEE Journal on Selected Topics in Signal Processing, vol. 11, no. 8, pp.1240-1253, 2017.