4V-08

中咽頭部収録音と LSTM-CTC を用いた咀嚼回数の自動推定

阿部 太樹[†1]　齊藤 隆仁[†2]　池田 大造[†2]　峰野 博史[†1]　西村 雅史[†1]

静岡大学情報学部[†1]　株式会社 NTT ドコモ[†2]

1. はじめに

我々は健康保持増進の観点から，咀嚼から嚥下に至る一連の食事行動を，中咽頭付近で収録した音情報を用いて簡便かつ詳細にモニタリングできるシステムの検討を行っている．今回特に咀嚼回数推定を目的とした検討を行なった．

先に安藤ら[1]は咀嚼，嚥下，発話をそれぞれ GMM(Gaussian Mixture Model)を用いてフレーム単位で識別した上で，食事行動と発話行動の認識を行う方法を提案した．ただ，GMM や，通常の DNN の学習には正確な時間情報ラベルが付与された大量のデータ（強ラベルデータ）が必要である．食事行動音に対してそれらを事後的に付与するのは容易ではなく，十分な性能が得られない一因となっていた．本研究では Connectionist Temporal Classification(CTC)を損失関数とする LSTM (Long Short-Term Memory) を咀嚼の推定に用いることで，時間情報ラベルのないデータでのモデル学習を可能とする．また，食事中にオンラインで弱ラベルを付与する手法によって，大量の学習データ収集を実現した．結果として限られた量の強ラベルデータで学習された従来方法と比較して，咀嚼回数の推定精度を大きく改善できる見通しを得たので報告する．

2. 提案手法

従来の DNN の学習では，入力系列の各時点において入力に対応する正解ラベル（以降，強ラベルと呼ぶ）が必要であった．CTC[2]は blank と呼ばれる空白ラベルを導入することで，長さの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り，イベントの種類，回数のみを含んだラベル（以降，弱ラベルと呼ぶ）で学習することが可能となる．松吉ら[3]は音響イベント検出において CTC を用いた弱ラベル学習の有効性について示している．CTC を用いることで大幅にラベリングコストが削減できるだけでなく，出力をイベント単位で得ることが可能となる．本研究でも LSTM-CTC を用いて咀嚼音データから咀嚼回数の推定を行うこととした．

Automatic Estimation of the Number of Chewings Using LSTM-CTC and Recordings around Oropharynx: Taiju Abe, Takato Saito, Daizo Ikeda, Hiroshi Mineno, Masafumi Nishimura

3. 実験

3.1. 実験機器・データ収集

ネックバンド型のフレームに取り付けた小型コンデンサーマイクを中咽頭部の皮膚に密着させることで食事中の咀嚼音を収録する．録音には IC レコーダを用いた（Linear PCM, 16bit 44.1KHz）．今回使用した実験用機材とその装着の様子を図 1 に示す．

図 1：収録用機材とマイク装着の様子

データ収集は 20 代の男子 8 名によって行った(約 8 時間)．被験者にはチューイングガムを摂取してもらい，同時に咀嚼回数をカウントしてもらうことで弱ラベルを作成した．また弱ラベルは咀嚼音の閉口音，開口音のそれぞれに「咀嚼」というラベルを付与した(実際の咀嚼回数は咀嚼ラベルの個数の半分となる)．本研究において咀嚼データに付与する弱ラベルと強ラベルの例を図 2 に示す(C:咀嚼)．

今回の弱ラベル作成において，食事中にオンラインでラベル付与を行うためにアプリケーションを作成した（図 3）．具体的には，被験者に咀嚼（＋嚥下）と同時にキーを押してもらうことでイベントログを作成し，そのログを弱ラベルとして用いることで弱ラベル作成のコストの削減を行った．今後のために嚥下のラベルも付与しているが，本報告では咀嚼音だけを対象として識別を行った．結果として，総咀嚼ラベル数 10200 個の咀嚼データを収集し，そのうち 5 名分(9400 個)を学習用データ，残りの 3 名分(800 個)を評価用データに用いた．また弱ラベルの 15 分の 1 程度の強ラベル(600 個)を人手で作成した．

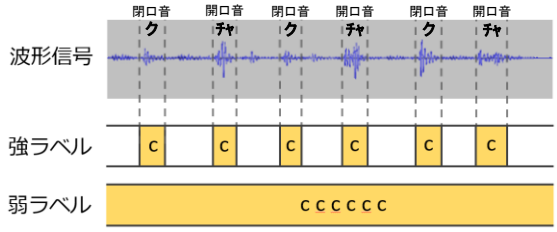

図 2：ラベル例

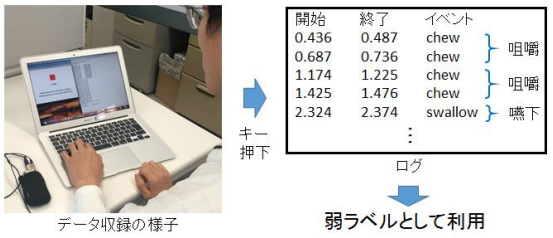

図 3：弱ラベル作成アプリケーション

3.2. ネットワーク・特徴量

LSTMの隠れ層のユニット数は200とした．パラメータの更新はミニバッチ数 50 で学習係数10^{-4}，1次モーメント係数0.9，2次モーメント係数0.999のAdamによって行った．特徴量は12次元のMFCC（Mel Frequency Cepstral Coefficients）に1次元のRMS（Root Mean Square）を加えた13次元とこれらの変化量Δと$\Delta\Delta$を加えた計 39 次元をフレーム幅 80ms，シフト幅 40ms で抽出を行い，LSTMの入力系列とした．

3.3. 評価尺度

咀嚼回数推定の評価尺度として MAPE（Mean Absolute Percentage Error）を用いた．MAPEは正解回数t_kと推定回数y_kで以下の式で算出する．

$$\mathrm{MAPE} = \frac{100}{N}\sum_{k=1}^{N}\left|\frac{y_k - t_k}{t_k}\right|$$

また LSTM の出力はフレーム単位での適合率，再現率，F値で評価し，LSTM-CTCはフレーム単位とイベント単位の適合率，再現率，F 値で評価を行う．イベント単位は正解ラベルと出力ラベルのイベントが重なっていた場合に検出とする．

3.4. 実験結果と考察

実験結果を表 1 に示す．またデータ量によるMAPE の推移を図 4 に示す．強ラベルと同程度の弱ラベルで学習した LSTM-CTC ではデータ量が足らず低い精度となっていたが，大量の弱ラベルで学習した LSTM-CTC では，限られた強ラベルで学習したLSTMよりも高い精度となった．CTCはスパイク状の事後確率が対応するイベントの中心付近に出現するという性質を持つため，フレーム単位では再現率が低くなっているが，イベント単位では正しく検出できていることがわかる．

表 1：実験結果

		フレーム単位			イベント単位		
使用モデル	MAPE (%)	適合率	再現率	F 値	適合率	再現率	F 値
LSTM	36.1	0.64	0.53	0.58	–	–	–
LSTM-CTC	10.8	0.81	0.41	0.54	0.89	0.90	0.89

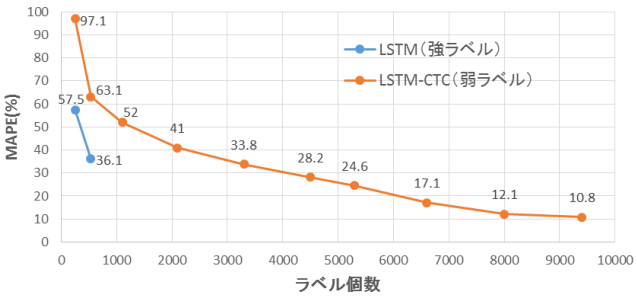

図 4：データ量と MAPE の推移

またガム以外の食物（グミ，スルメ，チョコレート，ポテトチップ）の咀嚼データを評価用の被験者(2 名)から追加で収集し，同様に LSTM-CTC での MAPE とイベント単位の評価を行った．結果を表 2 に示す．グミやスルメのような食感が似ている食物には同等の精度が見られた．またチョコレートのようなガムとは食感が異なる食物に対しても，わずかな推定精度の低下で抑えることができていた．

表 2：食物の違いによる評価

	MAPE (%)	適合率	再現率	F 値
ガム	10.8	0.89	0.90	0.89
グミ	11.5	0.86	0.90	0.88
スルメ	12.1	0.87	0.88	0.87
ポテトチップ	17.7	0.80	0.84	0.82
チョコレート	20.9	0.76	0.78	0.77

4. おわりに

弱ラベル付きの咀嚼音をオンライン収録する方法を提案し，大量の収録データで学習した LSTM-CTC を用いて咀嚼の検出を行い，従来法に比べて高い性能が得られる見通しを得た．今後は対象となる食物を増やすとともに，嚥下音の識別にも本手法を適用する予定である．

参考文献

[1] Jumpei Ando et al. NCSP 2018, pp.675-678, 2018.
[2] Alex Graves et al. ICML 2006, pp.369-376, 2006.
[3] Taiki Matsuyoshi et al. APSIPA 2018, pp.1918-1923, 2018.