咽喉音を利用した会話・摂食行動の認識

安藤 純平[†1] 齊藤 隆仁[†2] 川崎 仁嗣[†2] 片桐 雅二[†2] 池田 大造[†2]
峰野 博史[†3, †4] 西村 雅史[†3]

概要：「話すこと」「食べること」はそれぞれ嚥下機能，認知機能に深く関わり持つ．本研究では集音マイクと咽喉マイクの 2 チャンネルの音声から会話，摂食行動の検出を行う．提案法は GMM を使って会話・摂食に関連する詳細行動の識別をフレーム単位で行い，その識別結果を LSTM によって統合することで所望の会話・摂食行動を得る．実環境下において識別結果をヒューリスティックスに基づき時間軸上で平滑化する方法と比べて行動の検出精度を改善できることを確認した．

Conversational and Eating Behavior Recognition by Leveraging Throat Sound

JUMPEI ANDO[†1] TAKATO SAITO[†2] SATOSHI KAWASAKI[†2] MASAJI KATAGIRI[†2]
DAIZO IKEDA[†2] HIROSHI MINENO[†3, †4] MASAFUMI NISHIMURA[†3]

1. はじめに

超高齢社会が現実となり，高齢者の心身の状態，さらにはその家族や生活環境までを含めた関わりを幅広く正確に理解することで，高齢者の健康管理と介護サービスの効率化を実現したいという要望がある．その中でも特に高齢者の「食べること」は嚥下の能力に，「話すこと」は認知能力に関わる重要な行動要素である．我々はこれまで会議音声などの記録に用いる集音マイクに加え，ネックバンド型の咽喉マイク(図 1)から得られる２チャンネルの音響情報を用いて食事，会話行為に付随する詳細な行動の識別に関する研究を行ってきた[1-3]．ここでは発話や嚥下といった行動に関連する音区間を検出後に行動の識別を行なっていたが，実際の日常生活における多様で複雑な状況下では，このような音区間の検出自体が大変困難であるという課題があった．

本研究では，前処理としての音区間の検出は行わず，フレーム単位で行った識別結果の時系列データを事後的に統合することで，行動を検出する方法を検討する．このようなアプローチの場合，フレーム単位の識別結果を時間軸上で平滑化する処理が一般的に行われている．そのために，様々なヒューリスティック・ルールが用意されるが，多様な状況を想定してルールを書くのは容易ではない．一方，我々は２チャンネルの情報併用することで行動識別能力が高まることを確認しており，このような多元的情報も考慮したルールを用意するのはさらに困難であった．これらの複雑な状況に対処するため，本研究では RNN の一種である LSTM(Long Short-Term Memory)[4]を利用する．LSTM は長期依存の学習が可能なニューラルネットワークであり，フレーム単位の多元的な詳細識別結果を統合し，所望の粒度に近い形で行動（ここでは会話や食事などの行動）を出力できる可能性がある．本稿では，実環境下で収集したデータを用いて食事及び会話行動の検出を行うことで性能評価したので報告する．

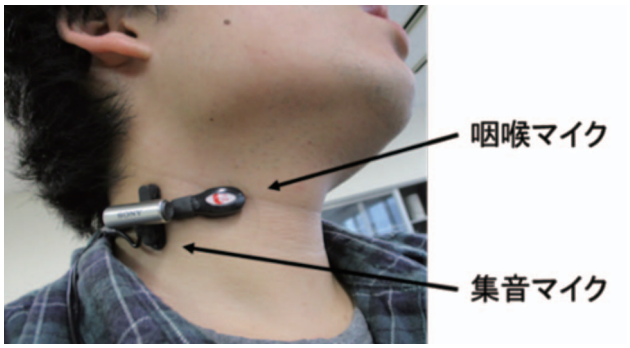

図 1 本研究で用いたマイクの構成とその装着例

2. 関連研究

日常生活行動を認識する研究は広く行われている．これまでの先行研究は，①室内環境に埋め込められたセンサにより利用者の行動を認識する方法[5-7]，②利用者の身体に装着したセンサを用いて行動を認識する方法[8-15]と，2 種類に大別できる．前者①は物品に埋め込むセンサが高コストである上に，予めセンサを設置した空間でしか行動を認識することができない．そのため，長時間，家庭内にいる人を除いて前者の方法を適用することは難しい．後者

†1 静岡大学大学院総合科学技術研究科
Graduate School of Integrated Science and Technology, Shizuoka University
†2 NTT ドコモ先進技術研究所　社会センシング研究グループ
Urban Sensing Research Group, Research Laboratories, NTT DOCOMO, INC.
†3 静岡大学学術院情報学領域
College of Informatics, Shizuoka University
†4 JST さきがけ / JST, PRESTO

Copyright (c) 2017 by the Information Processing Society of Japan

②は利用者の身体に装着したセンサを用いているので，入浴時間等を除いてほぼ丸々一日の行動を連続してセンシングできる．また，長時間センサを身体に装着することを想定しており，小型で装着負担の少ない加速度センサや集音マイクがよく活用される．そのため，前者①と比べて安価に運用できるために生活習慣や健康管理のようなライフログを目的とした行動認識では多く使われている．

ライフログのような長時間かつ多様な環境で収集したデータを対象とした行動識別の場合，限られた空間で収集したデータ比べて常に様々な種類の雑音が重畳している．とりわけ，音データの場合はその傾向が強い．そのため，予め時間軸上で対象行動の区間を検出し，その区間を切り出して識別することは難しい．とりわけ，対象行動の継続時間が異なる場合や実環境データに見られる複雑に行動が切り替わる場合はその境界時刻を自動的に判断することが難しいため正確に区間検出を行うのは困難を極める．我々が過去実施した研究[2]においても咽喉マイクに入る雑音や対象行動の継続時間差の影響を受けて区間検出の精度が低下している．また，精度向上を目的として DNN を用いて区間検出した場合についても雑音の影響を低減できた一方，継続時間の差に起因する検出誤りの改善は確認できなかった[3].

上記の手法に代わる方法として，前処理としての区間の検出は行わず，フレーム単位で行った識別結果の時系列データを事後的に統合することで，行動を検出する方法が存在する[8-15].この方法の利点として，雑音等を誤検出しても事後的に統合する過程で誤検出を排除できる．また，継続時間差や複雑に行動が切り変わる場合についても，短期間のフレーム単位で識別した結果を適切に統合することで対応可能である．しかしながら，適切な統合ルールを決めることは難しい．とりわけ，実環境データに見られる多様な状況を想定してルールを書くのは容易ではない．文献[13]では，ズボンのポケットに入れたスマートフォンと肩上に設置した集音マイクを使って 48 時間連続して加速度と周囲の環境音を収集し，LSTM を用いて食事や睡眠など 9 種類の識別を行っている．ここでは 1 分間の長いフレームで識別を行うために，フレーム内で最頻となった識別結果を最終結果とするルールを用いている．

その一方，本研究では食事中の会話など，会話・摂食の詳細な行動分析を行うことを目的としており，場合によってさらに短い単位での判別が必要になる．しかしながら，行動分析において所望の結果が得られると期待できるフレーム長を定めるのは難しい．また，文献[13]のような固定のフレーム長で統合する場合，行動の継続時間差や複雑に行動が切りわかる場合に関して柔軟に対応できない．そのため，心身状況を把握するという目的に合致するような詳細な行動情報を得ることは極めて難しいと考えられる．

また，文献[13]のように行動識別においてニューラルネットワークの活用する研究は多いが，その多くは学習データ構築を課題としている．文献[13]では，収集が容易な「睡眠」などと比べて「トイレ」「掃除」など収集が難しい対象行動の識別精度が低くなっている．他にも，文献[8]ではスマートフォンから得られる 3 軸加速度情報を基に利用者が「立っている」「座っている」「歩いている」「走っている」「階段を登っている/下っている」動作に関して 3 層パーセプトロンを用いて動作識別を行っている．しかしながら，こちらも学習データを大量に収集できなかった動作に関して決定木判別と比べて識別精度が低くなっている．その一方，学習データサイズの観点からニューラルネットワークを使わない行動識別も行われている．文献[14]では SVM を使って識別している．実運用時の所望からフレーム幅を 10 秒間としている．そのため，少量の学習データで識別をする必要があったので分類機に汎化性能の高い SVM を用いている．

3. 食事・会話行動の識別

3.1 識別対象の行動

本研究では様々ある日常生活上の行動の中から「食べること」「話すこと」に関連する行動として食事及び会話行動を対象とする．食事及び会話はそれぞれ嚥下の能力，認知能力と深い関係があることが知られている．嚥下の能力低下は誤嚥につながり肺炎の原因となることが知られている．加えて，健康的な食生活は身体の健康を保つのに欠かせない．食事量の減少による高齢者の低栄養は重篤な疾病をもたらす．同様に認知機能低下の兆候として会話の減少や独話の増加がある．更にはコミュニケーションの減少による高齢者のうつ病につながる．そのため，食事及び会話の行動を理解することで高齢者の健康維持管理の高度化が期待できるため対象とした．

3.2 概要

先行研究の知見と我々が過去行ってきた咽喉マイクと集音マイクを併用することの優位性を踏襲し，２つのマイクの信号から得られた２つの識別結果の時系列を事後的に LSTM で統合することで所望の行動を検出する方法を検討した．先行研究にあるようなフレーム単位の識別結果を時間軸上で統合する場合，フレーム単位の結果を時間軸上で平滑化する処理が一般的に行われている．そのために，様々なヒューリスティック・ルールが用意されているが，データや検出の目的が変更に応じてルールを決め直す必要が生じ，精度の高いルール発見には大変な手間がかかる．また，ライフログのような長時間かつ多様な状況で収集したデータに対してはフレーム単位の識別精度のばらつきが大きくなる傾向がある．そのため，すべての状況に対して

Copyright (c) 2017 by the Information Processing Society of Japan

対応できるルールを発見することは困難である．一方，我々は２チャンネルの情報併用することで行動識別能力が高まることを確認しており，このような多元的情報も考慮したルールを用意するのはさらに困難であった．これら複雑な状況に対処するため提案法では LSTM を用いて２つの識別結果の時系列情報（咽喉マイクの識別結果の時系列と集音マイクの識別結果の時系列）を統合する手法を検討した．LSTM は RNN の一種であり，長時間の相関を考慮可能なニューラルネットワークである．そのため，各情報源から得られた嚥下，咀嚼，発話などの多様で複雑な時系列生起パターンから，必要とする粒度で食事や会話を識別できる．

また，ニューラルネットワークを使った識別を行う場合，学習データの収集が課題となる．とりわけ，日常生活環境下で収集した多様な学習データにフレーム単位の詳細な行動ラベルを付与するのは大変困難である．提案法では，GMM であらかじめ識別した結果を LSTM で統合することで食事と会話を識別している．結果として，一度で音響的な特徴量から食事と会話を学習すると比べて，ニューラルネットワークの学習を効率化できる．そのため，学習データの収集のコストを下げられる．その一方で，食べながら話している」「テレビを見ながら食べている」といった複雑な行動の識別も学習によって可能になるものと期待している．

3.3 処理概要

提案する行動識別手法を図 2 に示す．入力は集音マイク及び咽喉マイクの２チャンネルの情報であり，独立に収録される．これらのデータはフレームの窓幅 25msec，フレーム周期 10msec で周波数分析され，それぞれ 39 次元の MFCC 特徴量時系列(速度，加速度成分も含む)に変換される．それら特徴量を用いて GMM によるフレーム単位の識別を行う．識別にニューラルネットワークの利用も検討したが，大量の学習データを事前に用意するのが難しい．そのため，先行研究[17,18]でも使われていて，少量の学習データでも過学習が起きにくいと考えられる GMM を選択した．集音マイク側については「発話」「雑音」「無音」の３つの状態の識別を事前に学習した GMM によって行う．一方，咽喉マイク側については，これら３つの状態に加えて，「嚥下」「咀嚼」の計５つの状態の識別を行う．なお，雑音については複数の種類のモデルを用意して識別性能の改善を図っているが，LSTM への入力としては１種類の「雑音」として扱っている．このようにして得られた２種類の詳細識別結果の時系列を LSTM に入力し，所望の粒度での行動識別結果を得る．集音マイクには周辺話者の発話を含む音声が記録される，一方で咽喉マイクは本人の発話音声だけが記録されるため，２つのマイクを利用することで，他者の発話と本人の発話を正確に区別することができる．その結果として会話と独話の識別にも繋がる．また，咽喉マイクは摂食行動の識別に有効である[16]．そのため，食事の識別も期待できる．なお，得られる結果はフレーム単位であり，連続性は保障されないが，この結果を平滑化処理するのは容易であると考えている．

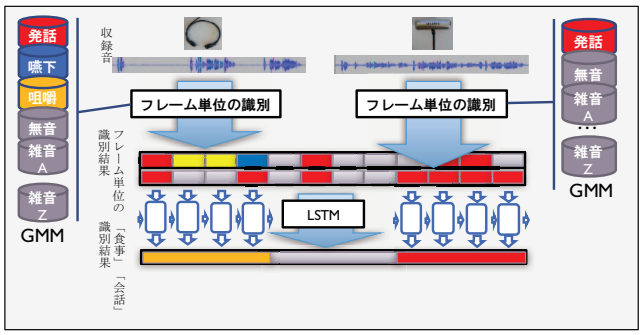

図 2　提案手法

3.4 GMM による詳細行動識別

GMM を用いた詳細行動のフレーム単位識別では，入力を咽喉マイク、集音マイクそれぞれの観測信号とし，フレームごとに抽出した特徴量からマイク毎に GMM で行動モデルを作成し，短時間のフレームごとに識別を行う GMM は，複数の正規分布を混合することで表されるモデルで，音を使った行動認識では広く使われている[17,18]．学習データサイズに応じて混合数を適切に調整することで少量の学習データを対象とした場合でも過学習を抑えられる．学習データ x の確率密度分布を $p(x)$としたとき，混合数 K の GMM M(x)は以下の式で表せる．

$$\mathrm{M}(\mathrm{x}) = \sum_{k=1}^{K} p(k) N(x; \mu_k, V_k)$$

3.5 ヒューリスティック・ルールによる識別結果統合

ここで，GMM を用いて識別したフレーム単位の識別結果をあらかじめ決められたヒューリスティック・ルールに基づき食事，会話行動の識別結果へ変換する方法について述べる．本研究で適用した変換規則は次のとおりである．

① 集音マイク側で発話と識別したフレームを会話に変換

② 咽喉マイク側の発話と識別したフレームは会話に，嚥下，咀嚼と識別した部分を食事に上書きして変換

③ 結果を時間軸上で平滑化

この規則は食事に関して咽喉マイク側の識別結果が適用される．一方，会話に関して，自発話は咽喉マイク側の結果を適用すること，他者発話を集音マイク側の結果を適用することで会話が検出できると期待できる．また，変換した結果を時間軸上で平滑化することで所望の粒度へ変換で

Copyright (c) 2017 by the Information Processing Society of Japan

きる．平滑化としては，文献[12,13]と同様に，60 秒間の区間毎にフレーム単位の識別結果が最頻となったものをその区間の識別結果とする処理を行った．

3.6 LSTM による識別結果統合

LSTM による識別結果統合では，前段で咽喉マイク，集音マイクそれぞれ GMM を用いてフレームごとに識別した２つの結果を１つの LSTM に入力して食事，会話行動の検出を行う．LSTM は Recurrent Neural Network(RNN)の拡張として提案されたモデルである．RNN の中間層のユニットを，過去の状態を記憶するメモリセルと，その記憶機関を制御するゲートからなる LSTM ブロックに置き換えることで，RNN の課題であった勾配消失・勾配爆発現象を抑えることができる．本研究では Forget Gate を導入した LSTM[17]を用いた．LSTM に入力する特徴量は，GMM による識別結果として咽喉マイク側は「嚥下」「咀嚼」「発話」「雑音」の 4 次元の one-hot ベクトル，集音マイク側は「発話」「雑音」の 2 次元の one-hot ベクトルが得られており，それらを連結し 6 次元のベクトルとしたものを使った．なお，GMM の識別で得られた「嚥下」「咀嚼」「発話」以外の結果は全て「雑音」として扱った．また，識別結果ではなく GMM のスコアを直接入力特徴量に使うことも検討したが，人手でラベル付けしたものを一部学習に使う半教師なし学習による精度改善を今後の実施することを目的として本研究では利用しなかった．中間層は 1 層で LSTM 層となっている．出力層は 3 つのノードを持ち，各ノードは食事，会話，および食事，会話以外の状態と対応しており，それぞれの行動，状態らしさを確率値で出力される．今回は，3 つのノードの確率値が最大となるノードの行動を最終的な出力とした．

4. データ収集

データは行動識別に使う GMM 及び識別結果の統合に使う LSTM の学習及びその評価のために 3 種類用意した．GMM の学習は 4.2 節の個別行動音及び 4.3 節の実験室行動音の 2 種類を使う．LSTM の学習には実験室行動音及び 4.4 節の実環境行動音を使う．なお，GMM による行動識別及び LSTM による識別結果の統合を評価するに際して，実環境行動音の一部，1 時間分を使った．

4.1 収録機器

3 種類のデータ収集は同一の機器を使って次に示す設定で行った．集音マイクは Sony 製 ECM-PC50 を使った．咽喉マイクは南豆無線電気製 SH12-iK を使った．収録は IC レコーダを使って実施した．IC レコーダは Sony 製 ICD-UX544F を使い，サンプリング周波数 44.1kHz / Linear PCM に設定した．また，ノイズカット等の機能は全てオフに設定した．

4.2 個別行動音

実験室内で個別に食事や会話をしている際の音を収集した．食事については３回分（食べ始めから終わりまでを 1 回と定義）収集した．食べ物は全てお弁当に統一している．データはのべ 20 分程度集まった．会話については 3 名が 5 分間の間，雑談する際の音声を収集した．データはのべ 60 分程度集まった．収集したデータは GMM の学習に利用する．

4.3 実験室行動音

実験室内にて食事や会話をしている際のデータで，12 名分，約 120 分間分用意した．収録は室内でお菓子や会話をしながらリラックスしている環境を模擬して行った．被験者に椅子に座ってもらい，スマートフォンを操作するなど普段の振る舞いをしてもらった．このデータはフレーム単位の識別に使う GMM の学習に用いる．そのため，嚥下，咀嚼，発話（自発話，他者の発話を含むすべての発話）の詳細な行動ラベル付けを行った．のべ 120 分間の音声データに付与した行動ラベルの内訳を表 1 に示す．咽喉マイク側には嚥下、咀嚼、およびマイク装着者自身の発話（自発話）のラベルを付与し、集音マイク側には自発話および周囲話者の発話すべてに対して発話ラベルを付与した。また，LSTM の学習に活用することを目的として別途、食事，会話行動のラベル付けも行った．のべ 120 分間のデータの内，食事は全体の 31%(約 37 分)，会話は全体の 21%(約 25 分)であった．食事ラベルは食事開始時から食事終了時まで付与した．途中，食事でない会話などの区間についてはラベルを外している．会話ラベルは一つの会話に対して，一度でも会話のやり取りがあれば会話ラベルを付与した．なお，食事及び会話ラベルはアノテータが音のみを聞いて判断して付与している．そのため，実際の被験者の行動と異なる場合も見られた．

表 1　詳細行動ラベルの内訳(実験室行動音)

ラベル名	ラベルの総数
嚥下	187
咀嚼	412
発話（咽喉マイク側）	527
発話（集音マイク側）	391

4.4 実環境行動音

実環境行動音は高齢者が 4 時間の間，家庭内を模擬した室内で食事を取りながら会話をしている際の音声を収録したものである．収録内容は 10 時から 14 時まで行い，その

Copyright (c) 2017 by the Information Processing Society of Japan

間はお菓子や飲み物を食べながら 4-6 名で雑談をしている様子が収録されている．また，途中に昼食をとっている．食事中の被験者同士の会話も記録されている．そのため，話しながら食事をするデータも含まれている．収録は家庭内音環境を模擬して行った．そのため，収録中にラジオを流したりして様々な家庭内の騒音を重畳させた．そのほか，エアコンの冷暖房のファンの音、ドアの開閉音など多種多様な背景音が重畳している．データ収録にはのべ 11 名の方に協力していただき、収録は合計 12 回行った。その結果，のべ 288 時間の音声が得られた。この収集したデータの中から昼食前後の音声を主に切り出して，のべ 8 時間程を利用した．切り出した音声には食事、会話のラベルを付与した。食事、会話ラベルは実験室行動音と同様の基準でラベル付けを行った。切り出した音声に付与した食事、会話のラベルの内訳を図 3 に示す。図 3 からわかるように，切り出したデータ（のべ 8 時間分）の 22%（約 1.8 時間）は会話，21%（約 1.7 時間）は食事となっている．また、切り出した音声のうち一部 1 時間に嚥下，咀嚼，発話の詳細な行動ラベルを付与した。ラベルの内訳を表 2 に示す．

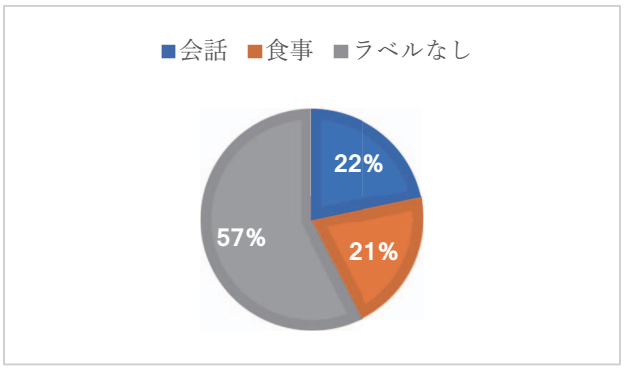

図 3　8 時間分のデータに付与した
食事，会話ラベルの内訳

表 2 詳細行動ラベルの内訳(実環境行動音)

ラベル名	ラベルの総数
嚥下	29
咀嚼	159
発話（咽喉マイク側）	233
発話（集音マイク側）	198

5.　評価実験

評価は 2 種類行う．初めに，5.1 節で嚥下，咀嚼，発話行動の識別性能を評価する．次に，5.2 節で LSTM による識別結果を統合する手法を用いた際の食事及び会話行動の検出性能を評価する．また，5.3 節では検出に対する LSTM の時間展開幅の影響を調査した．

5.1　詳細行動識別性能評価

実験は 4.2 節で言及した実験室行動音を学習に，実環境行動音の一部 1 時間程度を評価に使った．GMM はすべてのモデルを混合数 256 固定で作成し，ガウシアンの分散共分散行列は対角成分に制限した．評価指標は F-measure(再現率と適合率の調和平均)を用いた．評価はフレーム単位の識別結果に対して行った．咽喉マイク，集音マイクのそれぞれの結果を図 4，図 5 に示す．図 4 から，咽喉マイク側は嚥下，および咀嚼の識別精度が低いことがわかるが，嚥下に関しては評価データ中の絶対数が 29 回と少ないこと，および咽喉マイク装着者が動いたり首を動かしたりした際に生じる衣擦れ音を相対的に誤検出していため Recall に比べて Precision が低くなっている．

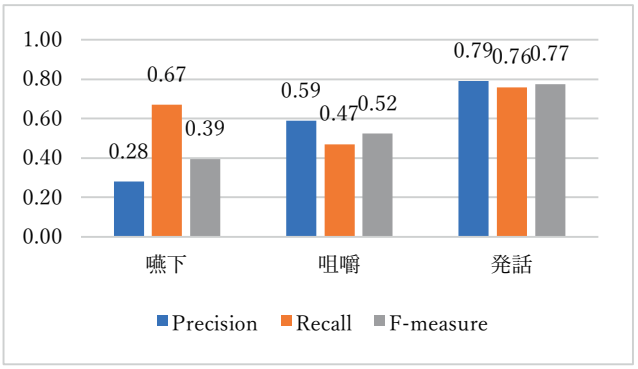

図 4　咽喉マイク側の識別結果

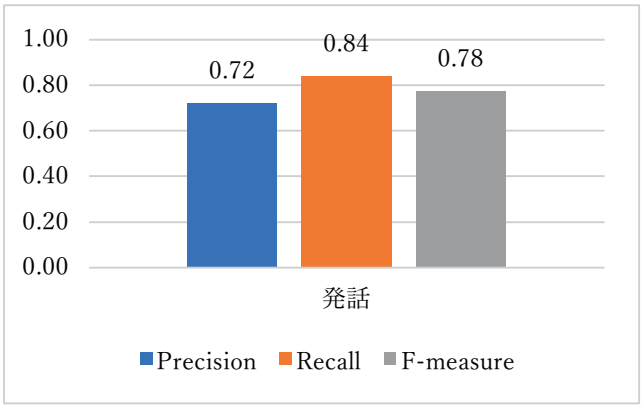

図 5　集音マイク側の識別結果

5.2　提案法性能評価

提案手法を用いて食事，会話の検出精度を評価する．学習データは実験室行動音と実環境行動音の 7 時間分を使った．評価には実環境行動音の 1 時間分を使った．評価指標は F-measure を用いた．LSTM の入力特徴量は各マイクの GMM を用いてフレームごとに嚥下，咀嚼，発話と識別した結果を使った．表 3 に LSTM の構造および学習パラ

Copyright (c) 2017 by the Information Processing Society of Japan

メータを示す．なお，比較対象として 3.5 節で言及したヒューリスティック・ルールに基づく食事，会話の検出も同様に実施して評価する．

図 6 にヒューリスティック・ルールに基づく方法，および提案手法の結果を示す．また，検出結果の一例として図 7 に評価に使った 1 時間分のデータの内，30 分間分切り出したデータに対して検出した結果を描画したものを示す．LSTM を用いた提案手法はヒューリスティック・ルールに基づく手法に比べて食事，会話ともに大幅に精度改善している．特に食事に関して大幅に識別精度が向上している．それは図 7 からも伺える．また，正解ラベルは大まかに付与したものであるために正解ラベルでは食事となっているが実際は食事しながら会話している区間が存在する．そういった食事中の会話を LSTM は実際の行動どおりに検出していた．この結果は大まかな正解ラベルで学習しても食事中の会話といった行動の切り替わりの激しい複雑な行動を識別できていたことを意味している．その一方で，より正確に正解ラベルを付与することでより正確に複雑な行動を識別できるかもしれない．

表 3　LSTM の構造と学習パラメータ

入力次元数	6
中間層	1
出力ユニット数	3
学習率	0.001
Drop out 率	0.2
活性化関数	ReLu[18]
最適化手法	Adam
時間展開幅	10 sec

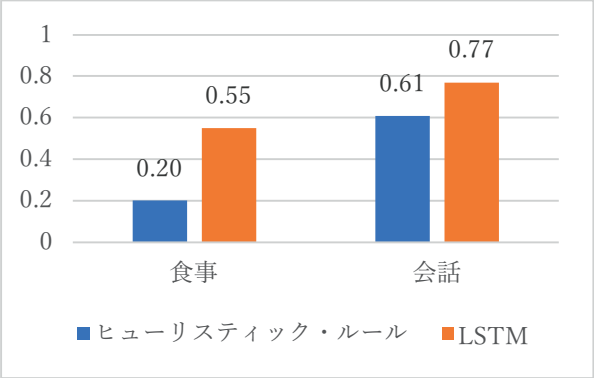

図 6　食事，会話の検出結果
- 既存手法（ヒューリスティック・ルール）と提案法（LSTM）の比較

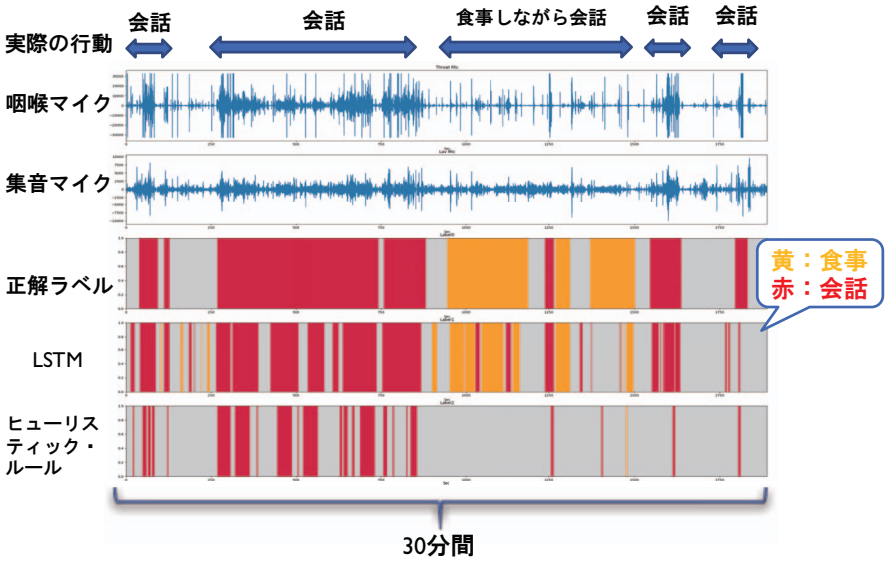

図 7　30 分間の実環境行動音に対する検出結果の一例
- 既存手法（ヒューリスティック・ルール）と提案法（LSTM）の比較

Copyright (c) 2017 by the Information Processing Society of Japan

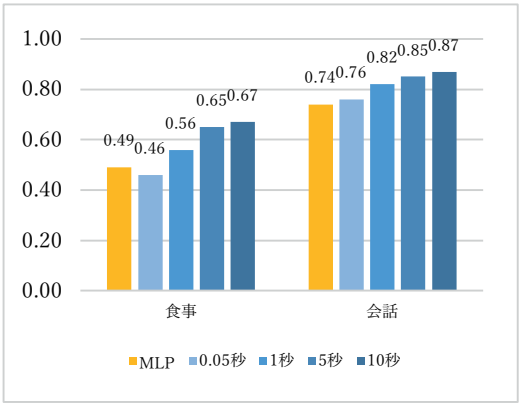

図 8　MLP 及び LSTM の時間展開幅別の識別結果（F-Measure）

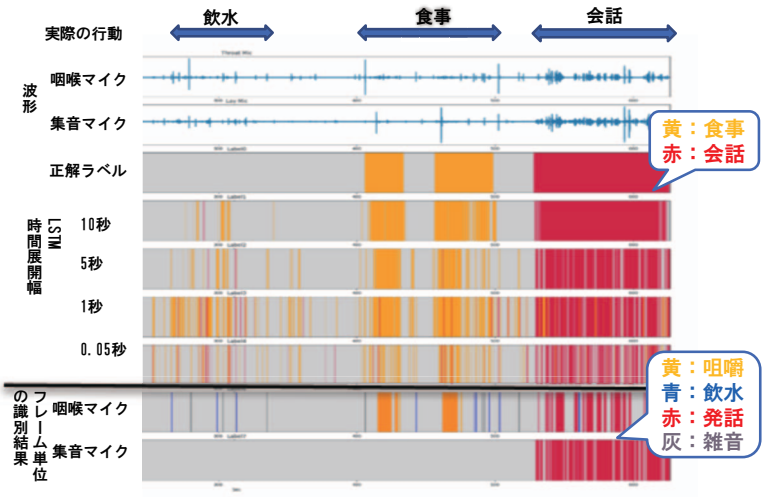

図 9　行動検出結果の一例 - LSTM 時間展開幅の影響

5.3 LSTM の時間展開幅性能調査

LSTM を用いて識別結果を統合するにあたり LSTM の時間展開幅の影響を調査した．この実験では LSTM の時間展開幅を 0.05 秒から 10 秒まで動かしてそれぞれの場合について LSTM を学習し，学習したネットワークを使って食事，会話の検出を行って，その精度をもとに時間展開幅について評価を行った．また，時間構造を使わないニューラルネットワークとして 3 層パーセプトロン(MLP)も同様に評価した．MLP および LSTM の学習，評価には実験室行動音を使った．そのうち 2/3 の 90 分間分を学習に，残りの 30 分間分を評価に使った．評価指標は F-measure を用いた．LSTM の構造と学習パラメータは今回検証する時間展開幅は除いて表 3 と同様である．なお，MLP については前後 2 フレームの結果も結合して入力に利用した．

図 8，図 9 に MLP および LSTM の時間展開幅別の検出結果とその一例(7 分間分)を示す．図 8 からわかるように，時間展開幅を長くなればなるほど検出の精度が向上していることがわかる．また本研究では，摂食行為を伴わない飲水に関しては食事ラベルを付与していないが図 9 を見るとわかるように時間展開幅を長くするにつれて誤検出が減少している．この結果は時間展開幅を広げることでフレーム単位の識別結果のパターンなど時間的コンテキストを学習していることを意味していると考えられる．また，飲水に別途正解ラベルを付与することで摂食行為を伴わない飲水と通常の食事に関して識別結果の時系列の違いを学習できる．その結果として飲水と通常の摂食を識別できるかもしれない．

5.4 考察

提案法の検出結果をみると会話に比べて食事の検出精度が低くなっている．また，ヒューリスティック・ルールによる手法に関しても同様に精度が低い．これは GMM による詳細行動識別の精度が原因であると考えられる．食事検出精度が低くなる現象は実環境行動音を評価対象とした場合に顕著に現れた．その理由として，実環境行動音の飲食物は多種多様あり，それに応じて咀嚼音も多様化していたことが考えられる．また，個人差もあり噛む速度や噛み方が少し変わるだけで識別できなくなる場合も見られた．それに対して，ポテトチップスのみを提供していた実験室行動音の咀嚼検出精度は実環境と比べて高いことを確認している．その結果として 5.3 節で示した最終的な食事検出精度も実環境行動音を使った場合と比べて精度が高くなっている．したがって，食事検出精度を改善するには実環境下における咀嚼や嚥下の識別精度向上が必要であると考えられる．また，咀嚼識別に関しては咽喉マイクの物理的な性質上必ずしも常に識別可能であるとはいえず，ある程度の誤りを許容するような仕組みが必要であると考えている．

また，5.2 節において，食事中の会話を実際の行動どおりに検出できたことを確認した．そこで得られた結果は大まかな正解ラベルで学習しても食事中の会話といった行動の切り替わりの激しい複雑な行動を識別できていたことを指している．本来は，食事中の会話といった複雑な行動を識別する場合は詳細な正解ラベルが必要になる．一方で，提案法は詳細な正解ラベルを付与することなく食事中の会話を識別できており，これは学習データ構築のために必要なラベル付けのコスト削減に繋がる．また，その反対により詳細な正解ラベルを付与することでより複雑な行動の識別を期待できる．ただし，目的やラベル付けの負担を考慮して慎重に検討する必要があるだろう．

Copyright (c) 2017 by the Information Processing Society of Japan

6. おわりに

本研究では嚥下、咀嚼、発話といった食事、会話行動に関連のある行動をフレーム単位で識別した結果に LSTM を適用して統合することで食事，会話行動を検出する手法を提案した．フレーム単位の識別結果をあらかじめ定めた規則に応じて変換するヒューリスティック・ルールに基づく手法と比べて大幅な精度向上を確認できた．

今後の課題として，より多様な状況を含むデータに対して評価することがあげられる．とりわけ，我々がユースケースとして考えている高齢者の健康管理では心身の状況と関係の深いより複雑な行動，例えば，会話について自分が積極的に話している，他者が積極的に話している，TV 等に話しかけている，を識別したいという需要がある．そのため，学習データ構築のコストを考慮しつつ今回できなかったより詳細なラベルを付与してより複雑な行動の識別ができるかどうかを分析する予定である．また，ニューラルネットワークの構造[19,20]を変更することで検出性能向上を図ることも検討している．

本研究では LSTM の用いた統合に関して食事検出精度に課題となった．その原因として食事に関連する詳細行動の識別精度が考えられる．2 章で取り上げた文献[14]では，行動の継続時間とフレーム幅の関係性について言及しており，我々はフレーム幅を対象の行動に応じて最適化することで識別に有効な特徴を抽出できると考えている．今後は識別精度の低かった嚥下，咀嚼を中心に，フレーム幅や特徴量などを最適化することで検出精度向上を目指す予定である．

謝辞

実験データを提供して頂いた静岡大学情報学部桐山先生，ご討論頂いた静岡大学情報学部西田先生，綱川先生に感謝します．

参考文献

[1] 西村雅史 他, "生体音と環境音の同時収録による高齢者の行動および身体状態識別に関する検討." 日本音響学会 2015 年春季研究発表会講演論文集, 2-4-9 pp.1309-1310 (2015)
[2] 安藤純平 他, "非侵襲簡易型身体状況認識システムに関する研究." 日本音響学会 2016 年春季研究発表会講演論文集, 1-4-4 (2016)
[3] 安藤純平 他, "身体状況認識システムにおける音イベント検出方法に関する検討." 日本音響学会 2016 年秋季研究発表会講演論文集, 2-6-2 (2016)
[4] Hochreiter, Sepp et.al., "Long short-term memory." Neural Computation, Vol.9, No.8, pp.1735-1780 (1997)
[5] 中川健一 他, "実社会指向アプローチによる認知症高齢者のための協調型介護支援システムの研究開発." 情報処理学会論文誌 , Vol.49, No.1, pp.2-10 (2008)
[6] Philipose, Matthai et.al., "Inferring activities from interactions with objects." IEEE pervasive computing, Vol.3, No.4, pp.50-57 (2004)
[7] Tapia, Emmanuel Munguia et.al., "Activity recognition in the home using simple and ubiquitous sensors." International Conference on Pervasive Computing, pp.158-175 (2004)
[8] Kwapisz, Jennifer R et.al., "Activity recognition using cell phone accelerometers." ACM SigKDD Explorations Newsletter, Vol.12, No.2, pp.74-82 (2011)
[9] Ravi, Nishkam et.al., "Activity recognition from accelerometer data." Aaai, Vol.5, No.2005, (2005)
[10] Bao, Ling et.al., "Activity recognition from user-annotated acceleration data." International Conference on Pervasive Computing, pp.1-17 (2004)
[11] Peng, Ya-Ti et.al., "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models." Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on. IEEE, pp.1218-1221 (2009)
[12] 玉森聡 他, "Recurrent Neural Network に基づく日常生活行動認識." 電子情報通信学会技術報告, Vol.116, No.189, pp.7-12 (2016)
[13] 玉森聡 他, "日常生活行動認識のための Recurrent Neural Network 構造の調査." 日本音響学会 2016 年春季研究発表会講演論文集, 2-4-1 pp.1-2 (2016)
[14] 大内一成 他, "携帯電話搭載センサによるリアルタイム生活行動認識." 情報処理学会論文誌, Vol.53, No.7, pp.1675-1686 (2012)
[15] Zhang, Zhengyou et.al., "Multi-sensory microphones for robust speech detection, enhancement and recognition." ICASSP2004, pp.781-784 (2004)
[16] Bi, Yin et.al., "Autodietary: A wearable acoustic sensor system for food intake recognition in daily life." IEEE Sensors Journal, Vol.16, No.3, pp.806-816 (2016)
[17] Ying, Dongwen et.al., "Voice activity detection based on an unsupervised learning framework." IEEE Transactions on Audio, Speech, and Language Processing, Vol.19, No.8, pp.2624-2633 (2011)
[18] Gers, Felix A et.al., "Learning to forget: Continual prediction with LSTM." Neural computation, Vol.12, No.10, pp.2451-1471 (2000)
[19] Nair, Vinod et.al., "Rectified linear units improve restricted boltzmann machines." Proceedings of the 27th international conference on machine learning (ICML-10), pp.807-814 (2010)
[20] Sutskever, Ilya et.al., "Sequence to sequence learning with neural networks." Advances inneural information processing systems., pp.1-9 (2014)
[21] LeCun, Yann et.al., "Neural machine translation by jointly learning to align and translate." Nature, Vol.521, No.7553, pp.436-444 (2015)

Copyright (c) 2017 by the Information Processing Society of Japan