J-024

DCNNを用いた聴診器音響センシングによるタッチアクション認識

Touch Action Recognition by Stethoscope Acoustic Sensing Using DCNN

増田 凪紗† 古川 広一‡ 矢入 郁子†‡
Nagisa Masuda Koichi Furukawa Ikuko Eguchi Yairi

1.はじめに

IoT機器は，我々消費者の家庭にとっても一般的となりつつある．総務省の情報通信白書によると，世界の IoT 機器数は 2018 年時点で 300 億台とされており，今後も高い成長率が予測されている．しかし，IoT 機器のユーザインタフェースは依然として制限されており，利用者に負担を強いる場合がある．例えば，連続操作や微調整をその都度音声対話で行うこと，機器数に比例したコントローラーを所有すること，機器表面の小さなボタンやスクリーンを操作することなどである．一方 HCI 研究では，タッチアクション認識の研究が進められており，手軽かつ直感的な操作を可能にする手段としてその重要性が高く評価され，注目を浴びている[1].

そこで本稿は，利用者の指が固体オブジェクトに接触したときに得られる音響情報を用いて，タッチアクションを識別するシステムを提案する．具体的には簡易的に改造した聴診器を使用して音響を取得するプロトタイプを実装し，Deep Convolutional Neural Network(以下，DCNN)，Long Short-Term Memory(以下 LSTM)，それらを組み合わせたモデル(以下，CNN-LSTM)を用いた 3 つのシステムによって，タッチアクションの識別を行う．

音響情報を用いたタッチアクション認識は，音の伝搬性質上，空気と比較して固体物体の方が遥かに効率的な点，利用者にセンサを認知されないよう配置可能な点の 2 つの利点が存在する．本稿で提案するシステムを用いることで，生活空間上に配置されている机や壁，ドアなどの固体物体をインターフェースとして，コンピュータとのインタラクションが可能となる．これらは健常者に加えて，視覚に不自由がある人や高齢により自由に移動ができない人にとっても簡易な入力を可能とする有用な技術となりうる．

2.関連研究

2.1.アクティブ音響センシング

アクティブ音響センシングとは，音響信号を送信し，その音響信号を受信することで状態を計測する手法である．アクティブ音響センシングのうち，重要な設計変数の 1 つは送信機と受信機の位置関係である．送信機と受信機が同一箇所の構成はモノスタティック構成と呼ばれ，そうでない場合はマルチスタティック構成と呼ばれる．モノスタティック構成では，ドップラー効果を利用した空間中のジェスチャ認識が行われている[2,3]．マルチスタティック構成では，物体の触れ方推定[4]やガラス表面上のタッチ位置推定[5,6]，骨伝導を用いた生体入力手法[7]，スマートフォンなどのモバイルデバイスの入力インタフェース拡張[8]などの研究が行われている．アクティブ音響センシングの特徴は，振動が発生しない微小なタッチアクションや，物体表面上の物体状態を高精度に認識可能な点が挙げられる．これらの利点は，音響信号を送信し，その音響信号を受信する際の変位を計測するというアクティブ音響センシングの性質のため，定常状態とその他の状態との識別が可能な点が精度の高い理由である．ただしアクティブ音響センシング手法は多くの制限があり，その内の重要な制限の 1 つにタッチの表現力が豊富ではない点が挙げられる．具体的には，パルスのように短時間で生じるタッチアクションによる音響情報などの識別が制限される．タッチの表現力を豊富にするために，パッシブ音響センシング手法による研究が行われている．次節でパッシブ音響センシングの関連研究について説明する．

2.2.パッシブ音響センシング

パッシブ音響センシングとは，外部から発せられる音響信号を受信する手法である．HCI 領域では，タッチアクションの際に生じる音響情報を受信し識別することでアクション推定を行う研究が多くされてきた．それらの研究は，生体を利用する研究，既存入力機器の入力範囲の拡張を目的とした研究，非入力機器を入力可能とする研究に大別することが可能である．

受信機器を身体に装着することで身体を入力インタフェースとする研究として，Hambone は圧電式の生体音響センサを手首に装着し，摘む動作や弾く動作の際に生じる音響をセンサが骨伝導により受信し，その音響から指の動きを識別する[9]．The sound of one hand はマイク式の生体音響センサを手首に装着し，指先のジェスチャの際に生じる音響をセンサが骨伝導により受信し，その音響から指先のジェスチャを識別する[10]．Skinput は，腕に生体音響センサを装着し，腕や指をタップした際に生じる音響から位置を識別する[11]．これらの生体を介した音響センシング手法は，入力に関して場所に制限されないことからいつでもどこでも利用可能な点が非常に利点である一方で，利用者に装着感を与えてしまう点や人体をセンサシステムに組み込むことで個人差やノイズに堅牢ではないという欠点も存在する．

既存入力機器の入力範囲の拡張を目的とした研究として，TapBack はスマートフォンの背面をタップする音響をスマートフォン内蔵マイクセンサにより受信し，タップ回数を識別する[12]．Toffee は機器端末の四隅に圧電センサを配置し，機器端末を卓上に隣接させ，卓上に与えられた外部刺激による音響信号が四隅のセンサに到着する音響到着時間差を計測することで，外部刺激の角度を識別する[13]．非入力機器を入力可能とする研究として，scratch input は質感のある物体表面をスクラッチした際に生じる音

†上智大学理工学部 情報理工学科 Faculty of Science and Technology, Sophia University
‡上智大学大学院理工学研究科 理工学部専攻情報学領域 Graduate School of Science and Technology, Sophia University

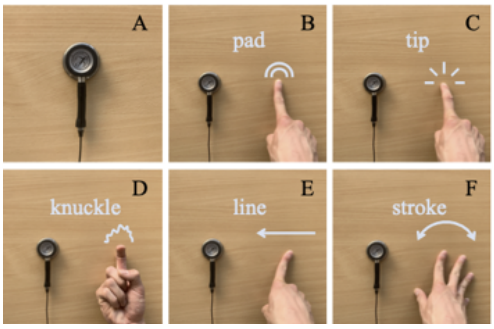

図 1　タッチアクション

図 2　実験環境概要

響を物体表面に取り付けた改造した聴診器で計測し識別する[14]．また，物体表面へのタップの音響からタップ種別を識別した研究もある[15,16]．これらの研究は，タップ種別を識別する設計変数として音響成分である振幅や位相，周波数成分が主に識別するための変数として利用されている．他には，タップ種別判定だけでなく，タップした音響からタップ位置推定を行う研究も存在する[17]．窓の四隅に圧電センサを配置し，窓表面をタップした際に生じる音響が圧電センサに到着する音響到着時間差を計測することで位置が特定される．これらの研究に共通する特徴は，利用者に何らかの装着感を与えずに自然な入力を可能とする点である．

3.提案手法とデータ収集方法

3.1.提案手法

本稿では，利用者の指と固体オブジェクトが接触したときに得られる音響情報からタッチアクションを識別する手法を提案する．対象とするタッチアクションは，指の腹，爪先，指の第 2 関節を用いたタップ 3種類(図 1-B, C, D)とスワイプ 1種類(図 1-E)，撫でるアクション 1種類(図 1-F)の 5 種類である．指の異なる部位のタッチアクションを採用した理由は，タッチする指の部位の構造の違いによる音響音分類の先行研究の存在にある[15]．先行研究では，Support Vector Machine を用いて分類を行なっているが，本稿では，DCNN，LSTM，CNN-LSTM を用いる．

3.2.提案システムの試作

音響センシング機器として，改造した聴診器を使用した(図 1 - A)．この機器は音響を取得する聴診器部と取得した音響をデジタル変換するマイク部の 2 つの構成から成る．聴診器は 3M 社の Littmann Classic Stethoscope を，マイクは Sony 社の ECM-SP10 を使用した．音響センシング機器は，縦 900mm，横 1800mm，厚さ 18mm の直方型の厚低圧メラミン化粧板テーブルの上部側中心に取り付けて使用した．マイクは USB オーディオ変換アダプタを介して PC との接続を行った．マイクのサンプリング周波数は 4100Hz とした．

3.3.実験参加者

指の異なる部位でのタッチアクション時の音響情報データを収集する実験に参加した実験参加者は，14 名（男性 11 名，女性 3 名）である．実験参加者の平均年齢は，22.5 歳，標準偏差は 1.79 である．実験参加者には，簡易なタップを行う実験であることと，実験所要時間が 30 分程度であることを伝え，参加を募った．また実験参加者への報酬は支払われなかった．

3.4.実験環境と実験手順

本実験で被験者がタッチアクションを行う対象である音響データ収集媒体は，厚低圧メラミン化粧板テーブルである．音響センシング機器は，媒体上部表面の中央に設置した．タッチアクション位置座標は，音響センシング機器と媒体の端との中心部，つまり媒体表面上の四分位置を指定した．図 2 に実験環境概要を示す．また，本実験中は常に空調機が稼働している環境下で行われた．実験参加者は椅子に座った状態で本実験に参加した．また計測時，実験参加者は，自身の指のみが媒体に接触するように実験監督者から指示された．実験参加者は監督者から実験内容の説明を受け，タップアクションを 10 回程度練習したのち，5 つのタップアクションをそれぞれ 50 回 x 2 セット，計 100 回ずつ行った．

4.タッチアクション識別モデルの構築

本節は，データセットの特徴量抽出方法，モデルに最適なデータセットの選択方法，使用するネットワーク構成の選出方法，堅牢性向上方法について説明する．

4.1.特徴量の抽出

3 節で採取したタッチアクションの音響情報を，1 つのタッチアクションの特徴量が 4000 次元となるようにローパスフィルターおよびダウンサンプリングを施した．さらに，高速フーリエ変換(以下、FFT)，メル周波数ケプストラム係数(以下、MFCC)を施すことで特徴量を算出した．FFT を採用した理由は，その性質上，畳み込みの計算を高速化できるためであり，DCNN の学習時間削減が見込めることを期待したからである．また，特徴量は 2000 次元とした．MFCC を採用した理由は，人間の聴覚特性を考慮することにより音声特性をできるだけ損なわずに特徴量の次元を少なくできるためである．MFCC の算出には，音楽およびオーディオ分析用ライブラリ LibROSA が使用された．また，特徴量は 24 次元と 64 次元の 2 つとした．

4.2.特徴量の選択

それぞれのモデルに，どの特徴量を用いたデータセットを入力するのが最適であるか調べ，一番識別率(F-score)が高かったものをデータセットとして採用した．データセットは，FFT，MFCC を施したもの，何も施さなかったもの

の 3 つを用いた．以降，それぞれのデータセットを FFT，MFCC（特徴量が 24 次元のものを MFCC24，64 次元のものを FMCC64），RAW と呼ぶ．

4.3.ネットワークの構成の選出

タッチアクションの分類には，周波数方向の関係性を学習できる DCNN，時間軸方向の関係性を学習できる LSTM，それらを組み合わせた CNN-LSTM の 3 つのモデルを使用した．DCNN のネットワークは車いすの加速度データを使用した路面状態推定の研究に基づき構成された [18]．LSTM と CNN-LSTM のネットワークは様々な層の組み合わせを試した上で一番識別率(F-score)が良かったものを採用した．これらのモデルの実装には深層学習ライブラリKeras，ハイパラメータのチューイングにはハイパラメータ自動最適化ライブラリ Optuna が使用された．

4.4.堅牢率向上手法

データ収集は，喧騒のない静かな環境下で行われたため，作成されたデータセットには雑音が含まれていない．しかし，実際にタッチアクションを行う環境は，必ずしも静かな場所とは限らない．よって，Data Augmentation を施したデータセットが作成された．具体的には RAW データに，ホワイトノイズ合成，音量調節，タイムストレッチを施した新たな 3 つのデータセットを用意し，それらを RAW データと合わせることでデータ量を 4 倍にした．そしてこれらをモデルに入力する際には，学習時間削減のために MFCC を施し，特徴量を 64 次元としたものを用いた．Data Augmentation には LiROSA が使用された．

4.5.モデルの評価方法

モデルを検証するために，14 人のデータセットで訓練したモデルを作成し，訓練用のデータセットで使用しなかった 1 名分のデータセットを評価データとして入力(1 個抜き交差検証(以下，LOSO))することで識別率が算出された．モデルの検証には機械学習ライブラリ Sklearn が使用された．モデルの評価には Accuracy と F-scoreを用いた．

5.結果

DCNN，LSTM，CNN-LSTM を用いたモデルにおいて，出力層は 5，最適化関数は Adam が採用されている．以下に各モデルの実装の詳細とその結果を説明する．

5.1. DCNN

DCNN のネットワークは，入力層，Conv1D - ReLU - Maxpooling1D(以下、MP) - Dropout からなる 4 つの畳み込み層，Flatten - Dense - ReLU - Dropout からなる全結合層，および Dense - softmax からなる出力層の 7 つの層で 構成された．全結合層のユニット数は 500 ，エポック数は 300，バッチサイズは 32，学習率は 5E-05 とそれぞれ設定された．DCNNは，RAW を用いると莫大な学習時間がかかり，MFCC を用いるとそれに合わせ，FFT を入力した時に構成したネットワークの構造を大きく変える必要があったため，FFT のみを用いた．識別率は，Accuracy 87.37[%]，F-score 86.71[%] であった．

5.2. LSTM

LSTM のネットワーク構成は．入力層，LSTM - Sigmoid - Dropout - BatchNormalization(以下、BN) , LSTM - BN からなる 2 つの隠れ層，Dense - softmax からなる出力層の 4 つの層で構成された．LSTM のユニット数は 80 と 20，バッチサイズは 40，学習率は 2E-03 と設定された．表 1 に，ネットワーク構造と F-score を示す．また，全てのパターンにおいて，入力データは MFCC64 が，入力層と出力層のユニット数とハイパラメータには同じ設定が用いられた．

5.3. CNN-LSTM

CNN-LSTM のネットワーク構成は，入力層，Conv1D - ReLU からなる畳み込み層，LSTM - Sigmoid からなる隠れ層，Dense - softmax からなる出力層の 4 つの層で構成された．LSTM のユニット数は 80，エポック数は 25，バッチサイズは 21，学習率は 5E-03 と設定された．表 2 に，ネットワーク構造とデータと用いた入力データ毎の F-score を示す．また，全てのパターンにおいて，入力データには MFCC64，FFT が，入力層と出力層のユニット数とハイパラメータには同じ設定が用いられた．識別率が高く出たデータセットは MFCC64 であった．

5.4. 考察

３つのネットワークのうち，DCNN が，Accuracy 87.37[%]，F-score 86.71[%] と最も良い数値となった．LSTM および CNN-LSTM は，それぞれ Acuuracy が 84.89[%]と 81.60[%]，F-score が 83.74[%]と 80.61[%] であっ

表1 LSTM ネットワーク構成の比較

Network Layout	F-score
LSTM - Sigmoid	83.06
LSTM - Sigmoid - BN	83.67
LSTM - BN - Sigmoid	83.17
LSTM-Sigmoid-BN-Dropout(0.5)	83.02
LSTM-Sigmoid-Dropout(0.5)-BN	83.38
LSTM-Sigmoid-Dropout(0.2)-BN	81.69
LSTM - Sigmoid - Dropout(0.5) - BN-LSTM(30) - Dropout(0.2) - BN	80.65
LSTM - Sigmoid - Dropout(0.2) - BN-LSTM(30) - BN	83.74

表2 CNN-LSTMネットワーク構成の比較

Network Layout	Data	
	MFCC64	**FFT**
Conv1D - LSTM	80.61	71.11
Conv1D - LSTM - Dropout(0.2) - BN	78.45	69.80
*(Conv1D - MP - Dropout(0.2))*2 - LSTM*	55.74	60.13
Conv1D - BN - Dropout(0.2) - LSTM - Dropout(0.2) - BN	78.71	69.11

た．DCNN は高い分類精度となったが，他のモデルと比べ学習に時間がかかった．学習時間は，DCNN が 69.41[sec]，LSTM が 9.41[sec]，CNN-LSTM が 11.09[sec]であった．学習時間は LOSO の 1 回分，つまり 1 人当たりの時間をもとに 14 人分の合計時間を求め， 1 回の平均で示されている．LSTM は表 1 に示したように隠れ層を重ねても識別率に大きく変化は見られなかった．また，通常はデータの特徴量が多いほど学習に時間がかかる傾向にあるが，MFCC の特徴量を増やしても学習時間への影響は少なかった．学習時間は，MFCC24 が 9.37[sec]，MFCC64 が 9.41[sec]であった．これは，本稿で使用した入力データが 12000 サンプルであり, そこまで膨大ではないためだと思われる．しかし，FFT，RAW それぞれの学習時間は 14.28[sec]，18.98[sec] であったため，特徴量が大きくなると学習時間は増える傾向にある．全てのモデルにおいて，Data Augmentation を施したデータセットを入力すると，精度が落ちた．これは，通常，Augmantation には wav データなどの音声データを用いるのに対し，本稿では，csv 化したデータを用いたことが原因だと考える．タッチアクションは，個体オブジェクトを叩く強度や，テンポに個人差が出やすい．また，入力にノイズが混じることも考えられる．そのため，実際にこのシステムを使用する際には，有用なデータセットであると考える．事前には，CNN と LSTM の欠点を補いあった CNN-LSTM の方が LSTM 単体よりも高い分類精度となることが期待されていたが，結果は LSTM 単体のほうが性能が勝る結果となった．これは，CNN と LSTM をスムーズにつなげるために，入力データを LSTM に適した形式で作ったことが原因ではないかと考えられる．入力データを CNN に最適な形式として畳み込み層に入力し，LSTM に入力するようなモデルを組めば，識別率が向上する可能性が残されている．

6.おわりに

本稿では，利用者の指が固体オブジェクトに接触したときに得られる音響情報を用いて，タッチアクションを識別するシステムを提案した．手法の検討のため，厚低圧メラミン化粧板テーブル，聴診器，オーディオインタフェースからなる音響センシング装置のプロトタイプを実装し，DCNN を用いたタッチアクションの識別精度が F-score で 86.71[%]となることを確認した．今後の課題として，CNN-LSTM のネットワーク構成をさらに工夫することで，識別精度の向上目指すことがあげられる．

謝辞

実験に参加してくださった方々に感謝します．本研究は科学研究費補助金，基盤研究 (B)17H01946 と基盤研究 (B) 20H04476 のもとで実施されました．

参考文献

1. F.B.Fernandez, J. Forrai, H. Hussmann, “Evaluation of User Interface Design and Input Methods for Applications on Mobile Touch Screen Devices”, *IFIP Conference on Human-Computer Interaction INTERACT 2009*, pp.243–246, (2009).
2. Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan, “SoundWave: using the doppler effect to sense gestures”, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’12), Association for Computing Machinery, New York, NY, USA, 1911–1914, (2012).
3. Kaustubh Kalgaonkar and Bhiksha Raj, ”One-handed gesture recognition using ultrasonic Doppler sonar”, In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’09), IEEE Computer Society, USA, 1889–1892, (2009).
4. Makoto Ono, Buntarou Shizuki, and Jiro Tanaka, “Touch & activate: adding interactivity to existing objects using active acoustic sensing”, In Proceedings of the 26th annual ACM symposium on User interface software and technology (UIST ’13), Association for Computing Machinery, New York, NY, USA, 31–40 ,(2013).
5. Collins, Tim, “Active acoustic touch interface”, Electronics Letters. 45. 1055 - 1056 , (2009).
6. Michael C. BrennerJames J. Fitzgibbon, “Surface acoustic wave touch panel system”, US Patent 4644100A, (1986).
7. Kentaro Takemura, Akihiro Ito, Jun Takamatsu, and Tsukasa Ogasawara, “Active bone-conducted sound sensing for wearable interfaces”, In Proceedings of the 24th annual ACM symposium adjunct on User interface software and technology (UIST ’11 Adjunct), Association for Computing Machinery, New York, NY, USA, 53–54, (2011).
8. Gierad Laput, Eric Brockmeyer, Moshe Mahler, Scott E. Hudson, and Chris Harrison, “Acoustruments: passive, acoustically-driven, interactive controls for handheld devices”, In ACM SIGGRAPH 2015 Emerging Technologies (SIGGRAPH ’15),Association for Computing Machinery, New York, NY, USA, Article 3, 1, (2015).
9. Travis Deyle, Szabolcs Palinko, Erika Shehan Poole, and Thad Starner, ”Hambone: A Bio-Acoustic Gesture Interface”, In Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers (ISWC ’07), IEEE Computer Society, USA, 1–8. , (2007).
10. Brian Amento, Will Hill, and Loren Terveen. “The sound of one hand: a wrist mounted bio-acoustic fingertip gesture interface”, In CHI ’02 Extended Abstracts on Human Factors in Computing Systems (CHI EA ’02), Association for Computing Machinery, New York, NY, USA, 724–725, (2002).
11. Chris Harrison, Desney Tan, and Dan Morris, “Skinput: appropriating the body as an input surface”, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’10), Association for Computing Machinery, New York, NY, USA, 453–462, (2010).
12. Simon Robinson, Nitendra Rajput, Matt Jones, Anupam Jain, Shrey Sahay, and Amit Nanavati, "TapBack: towards richer mobile interfaces in impoverished contexts”, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’11), Association for Computing Machinery, New York, NY, USA, 2733–2736, (2011).
13. Robert Xiao, Greg Lew, James Marsanico, Divya Hariharan, Scott Hudson, and Chris Harrison, “Toffee: enabling ad hoc, around-device interaction with acoustic time-of-arrival correlation”, In Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services (MobileHCI ’14), Association for Computing Machinery, New York, NY, USA, 67–76, (2014).
14. Chris Harrison and Scott E. Hudson, “Scratch input: creating large, inexpensive, unpowered and mobile finger input surfaces” In Proceedings of the 21st annual ACM symposium on User interface software and technology (UIST’08), Association for Computing Machinery, New York, NY, USA, 205–208, (2008).
15. C. Harrison, J. Schwarz, S. Hudsoni, “TapSense: Enhancing Finger Interaction on Touch Surfaces”, *UIST '11: Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp.627—636, (2011).
16. Pedro Lopes, Ricardo Jota, and Joaquim A. Jorge, “Augmenting touch interaction through acoustic sensing”, In Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces (ITS ’11), Association for Computing Machinery, New York, NY, USA, 53–56, (2011).
17. Joseph A. Paradiso, Che King Leo, Nisha Checka, and Kaijen Hsiao, “Passive acoustic knock tracking for interactive windows.”, In CHI ’02 Extended Abstracts on Human Factors in Computing Systems (CHI EA ’02), Association for Computing Machinery, New York, NY, USA, 732–733, (2002).
18. T. Watanabe, H. Takahashi, Y. Iwasawa, Y. Matsuo, I. Yairi, “Weakly Supervised Learning for Evaluating Road Surface Condition from Wheelchair Driving Data”, *MDPI Information*, 11(1), 2, (2020).