J-030

音声検索における身体的行動を利用した検索結果の適合性推定

山本　光穂† 　近藤　賢志† 　加藤　誠†† 　田中　克己††

†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
† デンソーアイティーラボラトリ　〒 150–0002 東京都渋谷区渋谷二丁目 15 番 1 号 渋谷クロスタワー 28 階
E-mail: †{miyamamoto,skondo}@d-itlab.co.jp, ††{kato,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　音声検索においてユーザの無意識的な身体行動を利用して検索結果の適合性を推定する手法について提案する．音声検索では I/F の制約から検索結果を操作する回数が制約されるため，タッチ操作等を利用して非適合文章を選択することは非現実的であり，その結果，適合フィードバックを行うことは困難である．そこで本論文では，音声検索における適合フィードバックの実現を目指し，ユーザの無意識的な身体行動を利用して検索結果に対する適合性推定を行う際に，どのような身体行動が有効であるのかを探索した．その結果，ユーザの無意識的な身体的行動を利用することで適合性推定が可能であること，また，顔の器官点の動きや表情の変化の情報が適合性推定に有効であることを示した．

キーワード　音声検索，ユーザ行動，暗黙的適合フィードバック

1. はじめに

音声検索はモバイル・デバイスの普及に伴い急速に普及している．音声検索は 2015 年現在 10 歳代のユーザの 55%，20 歳代のユーザの 41%が一日に一度以上利用している [8]．その結果，大手検索サービスの検索の約 25%が音声経由で実行されている [11]．また最近では，モバイル・デバイスだけではなく，パーソナルコンピュータ用のオペレーティングシステムにも音声検索が搭載されている [13]．このように音声検索は大変普及している．一方，音声検索に対する検索精度は例えば入力クエリが曖昧で短いという特徴等から検索品質に課題があると言われている [7]．

同課題に対する解決手法としては，音声検索に適合フィードバックを導入した上で検索結果に対する学習やクエリ修正等を行う手法が考えられる．しかし，音声検索は通常少ない画面操作，もしくは操作なしで目的とする検索結果を入手できる I/F が一般的であり，タッチ操作等を利用して非適合文章を選択することは非現実的である．その結果，適合フィードバックを行うことは困難である．そのため，音声検索で適合フィードバックを行う場合は，何らかのユーザ行動を利用した適合性推定が重要となる．その代替として考えられるのは，ジェスチャー等を使ってユーザがシステム側に検索結果に適合したか否かを明示的に伝達する方法である．同方法は，ユーザが明確に検索結果に対する適合性をシステムに伝えることができるため，適合フィードバックに対する高い効果が得られる．一方で，ユーザが操作を予め学習する必要がある，また，評価結果をシステムに伝えるためわざわざ何らかの行動をユーザが実行する必要がある等ユーザに負担が大きい．

そこで我々はユーザの無意識的な身体的行動に着目する．具体的には，音声検索中のユーザが無意識的に行う身体的行動，例えば表情の変化や視線移動等の行動を特徴量化し，それらを利用して検索結果に対する適合性を推定することが可能であれば，検索結果やクエリ等に対する適合フィードバックの実現が可能となるのではないかと考える．同仮説の検証のため，本研究では，まず，被験者が適合性フィードバックを意識せずに行った無意識的な身体的な行動のデータを獲得し，このデータを分析することで，どのような無意識的な操作が有効であったのかを探索した．その結果，ユーザの無意識的な身体的行動を利用することで適合性推定が可能であること，また，顔の器官点の動きや表情の変化の情報が適合性推定に有効であることを示した．

本論文の構成は次のようになっている．2 節では，検索タスクにおけるユーザの適合性推定手法やそれらの結果を利用した検索結果の向上手法に関連する論文を述べる．3 節では，本研究の課題を述べる．4 節では，実験システム及びデータセットを述べる．5 節では，提案した推定モデル及び同推定モデルを評価した結果を述べる．6 節では，今後の課題及び本論文の結論を述べる．

2. 関連研究

情報検索を行うユーザ行動のモデリングは，主にランキング学習やクエリ修正，対話的情報検索で用いられる．本節では，ユーザのモデリング手法とそのシステムへの応用という 2 つの側面から関連研究を述べる．

検索履歴を用いた適合性推定手法: 検索履歴からどのドキュメントをクリックしたか (クリック履歴)，また，どのドキュメントにどの程度滞在したか (滞在時間) という特徴量を利用して検索の支援を行う研究は数多く行われている．Joachims

らは，web 検索における検索結果のクリック履歴から検索結果の適合性を推定した上で，同適合性を検索結果のランキング学習の一つの特徴量として利用することで検索精度を向上させる手法を提案した [9]．White と Kelly らは検索結果の適合性をクリック履歴に加え，検索結果先でのドキュメントの滞在時間を利用して推定することで，検索精度を向上させる手法を提案した [18]．また，Dang や Wang らはクリック履歴や滞在時間を利用して，検索行動全体及びそれら検索行動中のそれぞれの行動，例えばクエリの入力や検索結果集合の表示，また，検索結果画面の表示に対する適合性を推定する手法を提案した [3], [17]．

身体的，生体的な情報を用いた適合性推定手法: 身体的や生体的な情報を用いてユーザ行動をモデリングし，それらの情報を利用して何らかのユーザ支援を試みる研究は，情報検索の研究領域にとどまらず様々な分野で行われている．Umemoto らは，検索中のユーザの視線を利用して検索意図とそれに基づく情報探索支援や，意図の推定，クエリ修正の手法等を提案した [15]．彼らの研究によると，視線から得られる各種特徴量を利用することでクリック履歴等の特徴量のみを利用するよりも高精度に適合性の推定が実現できることを証明した．Braunagel らは，ドライバの目や顔の向きから取得可能な情報を用いて，ドライバが注視しているタスクを推定する手法を提案した [2]．特に，彼らは EOG 信号を目から取り出し，同信号の波長の振幅を利用して記号化し特徴量として利用する手法を提案した．これらの特徴量と他の特徴量を組み合わせることで，運転中のドライバが実行しているタスク (ビデオ閲覧/読書/電子メール閲覧/なし) の分類を高精度に実現できることを示した．

Hassan らは音声検索における入力クエリから得られる音響的，言語的等の情報を利用してクエリ修正を実施する手法を提案した [6]．彼らの研究によると，特に音響的な特徴，例えば発話速度や発話の強さ等がクエリ修正の推定精度向上に一定程度寄与することを示した．

ユーザの感情を推定した上で，それらをユーザモデルの一つの特徴量として利用することにより，検索精度を向上させる研究も多く行われている．Feild らは，検索過程におけるユーザの不平・苛立ち等をクエリログとセンサ情報等を利用して推定する手法を提案した [5]．これは，検索結果の満足性に加え検索過程の最適性を追求するという考えに基づく．同提案は，検索過程における不満を 6 種類定義し，それらを予測することでより最適な検索結果を提示する手法である．

また，生体信号を用い適合性を判断する手法等も提案された．Eugster らは計測した脳波に基づき検索トピックの単語の適合性を推定した [4]．彼らの研究によると，特に事象関連電位が単語の適合性推定に有効であることを示した．

最後に，これらユーザの行動を複合的に用いてユーザの検索意図を推定する手法を紹介する．Moshfeghi らは検索時のユーザの感情，身体的，及び検索行動を複合的に利用してユーザの検索意図を推定する手法を提案した．彼らはビデオ検索システムを利用するユーザの生体情報を含むユーザの様々な行動を特徴量化することで，検索意図を分類できることを示した [12]．我々の研究の特徴は，彼らの研究は検索意図の分類のみにとどまっているのに対して，我々の研究は個々の検索ドキュメントの適合性の推定や不適合の理由の推定手法までモデル化し検証を行っている点等が挙げられる．

3. 課題及び検証手法

本研究の目的は，音声検索に適合フィードバックを導入する際に，有効な身体的特徴を検証することである．本目的に対応する $RQ1$ を以下と定義する．

[$RQ1$] 音声検索においてユーザの行動を観察することによって，ユーザに提示した情報の適合性の推定ができるか．また，どの行動が適合性の推定に寄与するか．

また，推定モデルを用いて推定した適合性の用途は，クエリ修正や検索結果の最適化を想定している．クエリ修正については，システム側が音声認識で認識したクエリがユーザの意図したものか否か，また，検索結果の最適化については，どのドキュメントにユーザの検索意図が適合しているか推定できることが重要である．同目的に対応する $RQ2$,$RQ3$ を以下と定義する．

[$RQ2$] 提示した情報に対してユーザが不適合とみなした場合，ユーザの行動から不適合の理由が音声検索の間違いに起因するものか，それとも検索意図に適合しない事に起因するかを推定できるか．また，どの行動が推定に寄与するか．

[$RQ3$] 提示した情報に対してユーザが適合とみなした場合，ユーザの行動から適合するドキュメントが具体的にどれであるかを推定可能か．また，どの行動が推定に寄与するか．

同 RQ を検証するために，1) 本研究ではまず音声検索行動時のユーザの行動を記録する実験システム，及び実験システムで利用するデータセットを構築した．2) 次に，被験者に開発した実験システムを利用させる事によって，(a) ユーザが適合するドキュメントを見つけることができた場合 (**S**uccess: **S**)，(b) 適合するドキュメントをユーザが見つけることができず，その原因がユーザの検索意図との相違に起因する場合 (**I**ntent **E**stimation **E**rror: **IEE**)，(c) 適合するドキュメントをユーザが見つけることができず，その原因が音声検索のエラーに起因する場合 (**V**oice **E**rror: **VE**)，それぞれのユーザの行動 U_S，U_{IEE}，U_{VE} を記録した．3) 次に，記録した各ユーザ行動 U_S，U_{IEE}，U_{VE} を利用し，各特徴量 F_S，F_{IEE}，F_{VE} を導出した．4) 次に，同特徴量を利用してそれぞれの RQ に対

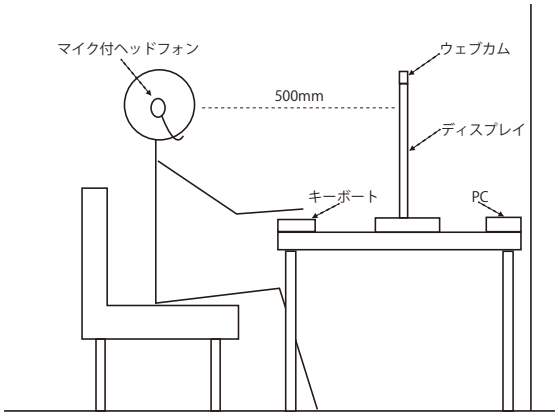

図 1　実験システムの構成

応する推定モデル，具体的には，ユーザが提示した検索結果に対して適合していると判断しているか否かを推定する推定モデル $P_{S/E}$ (**S**uccess or **E**rror: **S/E**)，ユーザが提示した検索結果に対して不適合とした理由を推定する推定モデル $P_{I/S}$ (**I**ntent or **S**peech: **I/S**)，ユーザがどの検索結果に対して適合するかを推定する推定モデル P_{RD} (**R**elevant **D**ocument: **RD**) を構築した上で，各推定モデルの推定性能を考察することにより，各 RQ の検証を行った．

4. ユーザ行動取得のための実験

本節では実験システム，実験システムにおける音声検索フロー，データセット，プロトコルを説明する．

4.1 実験システム

実験システムの構成を図 1 に示す．実験システムは音声検索の機能，及び音声検索時のユーザの行動を記録する機能を有する．実験システムは，検索結果を提示するディスプレイ，検索結果の読み上げによる提示やユーザの発話を記録するマイク付きヘッドフォン，ユーザが検索結果を選択するためのキーボート，被験者の検索行動を動画にて記録するためのウェブカメラ，及び PC で構成する．実験システムによって記録するユーザの行動は，ウェブカメラによって被験者の上半身を撮影した顔画像 (640x480，20fps)，マイク付きヘッドフォンで録画したユーザの発話 (44.1kHz/16Hz，wav データ)，及びキーボートの操作履歴である．被験者の顔画像は，ユーザの顔表情や顔の動き，頭の動きや視線等を捉えることを目的とする．ユーザの発話はユーザの発話の内容や発話のタイミング等を捉えることを目的とする．キーボートの操作履歴は検索結果の滞在時間や，どの検索結果を適合とみなしたか，等を取得する際に利用する．同データを利用したユーザの行動の特徴量化の手法は 5.1 節で説明する．実験時においては，ユーザの顔がディスプレイから 500mm の位置に来るようにユーザの着座する場所をその都度調整する．

なお，実験システムのハードウェアはウェブカメラとマイクロフォン及びキーボートでありユーザの行動履歴を取得する際に利用する．出力結果は画面と音声両方で出力する構成になっている．このような構成と同様の構成を取りうる音声検索のハードウェアは，ノート PC が最も近い．また，スマートフォンやカーナビゲーションを装着した車等も同様のハードウェア構成を取る．ただし，スマートフォン，車共にディスプレイが今回の実験システムと比較して小さいため，提示できる検索結果数や文字数に制約が生じる．また，モバイル環境下での利用が想定されるため，安定してウェブカメラでのユーザの顔画像が取ることが期待できない等の問題がある．

4.2 実験システムにおける音声検索フロー

本節では，本実験システムを利用してユーザに課す音声検索タスクのフローを説明する．図 2 に本実験システムとユーザとの音声検索フローの例を示す．実験システムはユーザに対して検索意図 i，及びその際にユーザに発話させる初期クエリ q_1 をディスプレイを通じて提示する．ユーザは指定された初期クエリ q_1 を発話する．実験システムはユーザの初期クエリ q_1 の 2 秒後に D_1 を D_{FRD}，D_{DI}，D_{VE} 中から実験条件に基づき選定した上で画面及び音声経由で提示する．なお，D_{FRD} は検索意図に適合する検索ドキュメント集合，D_{DI} は検索意図に適合しない検索ドキュメント集合，D_{VE} は音声認識エラーが発生した際に検索システムから返される事を想定した検索ドキュメント集合を示す．同検索ドキュメントの詳細は 4.3 節で説明する．音声読み上げの終了後，適合するドキュメントが検索結果 D_1 に含まれるとユーザが判断した場合，ユーザは適合するドキュメントの番号 n に相当するキーボートを押下する．この場合タスクは終了し，次の実験に移行する．一方で，適合する検索結果が D_1 に含まれないとユーザが判断した場合，修正クエリ q_2 を発話する．q_1 は予め指定されたクエリをユーザに発話させるのに対して，q_2 はユーザ自身が提示された検索意図 i に合致するクエリを自由に発話する事を許す．実験システムはユーザからのクエリの入力後，必ず q_2 に応じて D_2 として検索意図に適合するドキュメント集合である D_{FRD} を返す．ユーザは再び検索意図に適合する情報が D_2 に含まれると判断した場合，適合するドキュメントの番号 n を選択する．ユーザが適合する D_2 に含まれないと判断した場合，ユーザは再度修正クエリ q_3 をユーザに発話させる．以上で一回の実験タスクが終了する．

音声検索部の動作例を図 2 に示す．この例では，ユーザに i =“スターウォーズの内容について深く知りたい” を検索意図として，q_1 =“スターウォーズ” を初期クエリとして与えている．音声検索部は検索結果 D_1 として検索意図に適合しない検索結果 D_{DI} =“スターウォーズの上映映画館” を提示している．その結果，ユーザは検索結果 D_{DI} が検索意図に適合しないと判断している．その結果，ユーザはクエリ q_2 = “スターウォーズ　内容” とクエリを再投入しており，その結

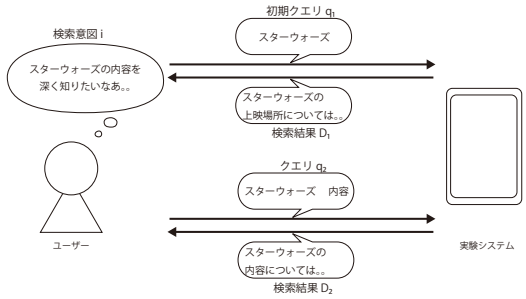

図 2　実験システムとユーザとのインラクション

果音声検索部は検索結果 D_2 として検索意図に適合する検索結果 D_{FRD} =“スターウォーズの内容” を提示している．

実験システムがユーザに提示する検索結果の件数は 6 件であり，ディスプレイには音声認識結果及びそれぞれのドキュメントのタイトル，URL，スニペット (100 文字) を提示する．音声読み上げにおける検索結果は，音声認識結果及びそれぞれのタイトルを読み上げる．ユーザの行動を記録する期間は検索結果 D_1 を提示してから検索結果 D_2 を提示するまでの期間である．

4.3　データセット

本実験で利用したデータセット T は検索意図 i，初期クエリ q_1，ドキュメント集合 D_{FRD}，D_{DI}，D_{VE}，クエリの音声認識エラーワード q_e によって構成する．これらデータセット集合を NTCIR INTENT-1 データセット及び Experimental data for Optimizing Search Result Presentation を利用して生成した [14]，[10]．NTCIR INTENT-1 のデータセットは 与えられたクエリに対して，ランク付けされた検索意図がサブトピック文字列として定義されている (注1)．また，Experimental data for Optimizing Search Result Presentation は NTCIR INTENT-1 データセットにおける検索意図の相関を定量化したデータセットである (注2)．

データセットの作成手順は以下の通りである．

（1） 初期クエリ q_1 を NTCIR INTENT-1 データセットから選定する．選定手法は NTCIR INTENT-1 のデータセットに含まれる検索クエリ (100 件) を利用して検索エンジンに検索クエリとして投入し，検索結果が多かった検索クエリ上位 30 件を利用する．

（2） D_{FRD}

（a） 選定した初期クエリ q_1 に対して NTCIR INTENT-1 データセットのサブトピック文字列のスコアを利用して，最も値が高いものを検索意図 i として選定する．

（b） 選定した検索意図をクエリとして，Bing Search API(注3) を利用することにより web ドキュメントを 20 件入手する．

（c） 入手した web ドキュメントに対してクラウドソーシングを利用して選定した検索意図 i に適合するかを 5 段階で評価 (1web ドキュメントあたりワーカー 10 名で評価)．その上で同評価結果の平均値に基づきデータを評価値が高い 6 件を選定する．

（3） D_{DI}

（a） D_{FRD} の選定時に選定した検索意図 i に対して最も検索意図に相関がない検索意図 i_{DI} を Experimental data for Optimizing Search Result Presentation から選定する．

（b） 選定した検索意図 i_{DI} をクエリとして Bing Search API を利用して web ドキュメントを 20 件入手する．

（c） 入手した web ドキュメントに対してクラウドソーシングを利用して選定した検索意図 i に適合するかを 5 段階で評価 (1web ドキュメントあたりワーカー 10 名で評価)．その上で同評価結果に基づき，評価値が低い 6 件を選定する．

（4） D_{VE} 及び q_e

（a） 初期クエリ q_1 を音声合成エンジン rospeex(注4) で合成して，その後同合成結果を Julius(注5) にて認識させ，認識結果の N-Best 解を入手する．その上で認識結果中に含まれる認識エラークエリ q_e を初期クエリ q_1 と比較することで選定する．

（b） 選定した q_e をクエリとして Bing Search API を利用して web ドキュメントを 6 件入手．同ドキュメントを D_{VE} とする．

4.4　プロトコル

4.1〜4.3 で説明した実験システム，タスクフロー，データセットに基づき，ユーザ実験を実施することによって音声検索実行時のユーザ行動データを収集した．被験者は 15 名 (男性 10 名，女性 5 名)，年齢は 18 歳から 24 歳までであり，すべての被験者は日本語を第一言語とする．実験時の報酬は 2000 円である．実験時間は一時間半である．

各被験者に対しては，まずユーザ実験の内容及び趣旨を説明した．次に実験システムの利用方法の説明を行った．ただし被験者には同実験における検索結果は予め検索結果が決められているわけではなく，その都度音声認識を行った上で検索結果をその都度算出していると説明した (注6)．次に，今回

(注1)：http://research.nii.ac.jp/ntcir/permission/ntcir-9/perm-ja-INTENT.html から入手可能

(注2)：http://www.mpkato.net/datasets/から入手可能

(注3)：http://datamarket.azure.com/dataset/bing/search

(注4)：http://rospeex.org/top-ja/

(注5)：http://julius.osdn.jp/

(注6)：実験後，被験者に対して「本システムにおける検索結果が予め決められていること 対して気がついたか否か」と質問したところ，15 名中 8 名が

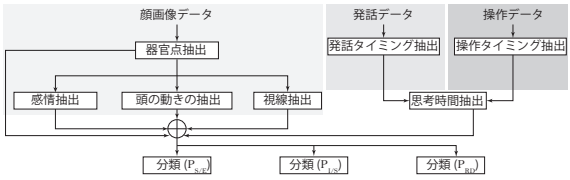

図 3　本研究における適合性推定モデル

図 4　各期間の定義

作成したデータセット以外での練習用データセットを用いて検索タスクを 5 回練習した後，本番のユーザ実験として 30 回検索タスクを実施させた．

今回の実験における検索タスクはクエリの種類 30 種類 に対して検索結果データ集合が 3 種類あることから，計 90 種類の検索タスクの種類が存在する．実験順序や検索タスクの分配はラテン方格法に基づき決定し，各被験者に実験を割り振った．

以上の条件下で実験を実施した．実験結果 450 件のユーザの検索行動 U を記録した．

5. 解　　析

本節では，ユーザの行動ログとして記録したデータの特徴量化手法を述べる．次にそれら特徴量を利用した分類手法及び分類結果を述べる．最後に同結果を考察する．

5.1 特 徴 量 化

以下に各々の特徴量の内容及び生成方法を説明する．

Thinking Time:(TT)　は検索結果を提示してからユーザの反応までかかる時間を特徴量化したデータである．具体的にはクエリ q_2 が投入されるまでの時間，もしくはユーザが適合するドキュメントを発見し選択するまでの時間である．同データを特徴量として利用することで，提示された検索結果を理解する時間や，新たにクエリを投入する際にユーザがクエリを考える時間等が与える影響などを推定モデルに反映することが可能であると考える．同データはユーザの操作履歴及び発話データを解析することにより生成した．

EMotion:(EM)　は顔の感情の変化を特徴量化したデータである．同情報を特徴量に利用した理由は，適合しない情報が提示された場合などに不満そうな表情をする，また適合した情報を提示した際は満足した表情をする等，ユーザの行動等の入手を期待するからである．動画データに基づきユーザの音声検索中の感情を Neutral/Happy/Surprised/Puzzled/Disgusted/Afraid/Sad の 6 種類に推定した上で，特定の期間における検索行動中の各感情の割合を算出した．感情の推定には，Insight SDK(注7) を利用した．

Eye Gaze:(EG)　は視線情報から得られる情報を元に特徴量化したデータである．同データを特徴量とすることにより，ユーザが画面を見ているか否か，もしくは検索結果候補の中からどの検索結果に注目しているか，等を捉えることが可能であると考えられる [16]．同特徴量は，Insight SDK を利用して動画データから視線を抽出した上で，それら視線の動きの移動量 $[f_9]$, 座標値 $[f_{10}]$,$[f_{11}]$ を導出した．

Facial Points:(FP)　は顔の器官点の動きを特徴量化したデータである．具体的には，顔のパーツ 68 点 (輪郭 17 点，眉 10 点，目 12 点，鼻 9 点，口 20 点) の器官点を取ったものである．同特徴量は顔のパーツのそれぞれの検索行動中の動きを捉えることができる．これによって，例えば目の動きから目を閉じているか否か，また口を動かしているか等の情報を得ることができる．同データの抽出は CLM-Franework(注8) を利用して動画データから顔の特徴点の座標を抽出・正規化した後，各器官点の移動量を区間の時間で割ったものを特徴量 $[f_{12}$-$f_{91}]$ とした．

Head Pose:(HP)　は頭の動きを特徴量化したデータである．同データを特徴量として利用することで，検索意図に適合しない情報が提示された場合に首をかしげる，また，適合するドキュメントを探索している際の首の上下の動き等のユーザの行動を捉える事が可能であると思われる．同特徴量は，Insight SDK を利用して動画データに基づき抽出した．

以上に示した 5 つのカテゴリから表 5.1 に示す種々の特徴量を計算することにより，各 RQ に対応する推定モデルを構築する (図 3)．ユーザの行動ログの特徴量化に際しては，図 4 に示すように，3 つの区間にユーザの行動を区切った上で，それぞれの区間のユーザ行動ログを利用して特徴量化する事とする．

Session 1　画面に検索結果を提示，読み上げが終了するまでの期間．

Session 2　検索結果の読み上げが終了してから，ユーザが何らかの行動，例えば検索結果を選択する，また，検索クエリ等を入力する，等の行動を起こすまでの期間．

Session 3　ユーザが何らかの行動を起こしてから次の画面が提示されるまでの期間．

「実験結果が予め決められている事に対して最後まで気が付かなかった」と回答した．

(注7)：http://sightcorp.com/insight/

(注8)：https://github.com/TadasBaltrusaitis/OpenFace

表 1　本研究で利用する音声検索時のユーザの行動特徴量一覧

TT	操作履歴
$[f_1]$	ユーザが検索結果に基づき何らかの行動を起こすまでの時間
EM	感情
$[f_2]$	ユーザの表情が “無表情” であった時間の割合
$[f_3]$	“幸せ”
$[f_4]$	“驚き”
$[f_5]$	“怒り”
$[f_6]$	“うんざり”
$[f_7]$	“恐れ”
$[f_8]$	“悲しみ”
EG	視線
$[f_9]$	視線の移動量
$[f_{10}]$	視線の x 座標平均
$[f_{11}]$	視線の y 座標平均
FP	顔の器官点
$[f_{12} - f_{39}]$	輪郭の移動量
$[f_{40} - f_{49}]$	眉の移動量
$[f_{50} - f_{61}]$	目の移動量
$[f_{62} - f_{70}]$	鼻の移動量
$[f_{71} - f_{91}]$	口の移動量
HP	頭部の移動
$[f_{92}]$	上下の回転量の平均
$[f_{93}]$	左右の回転量の平均
$[f_{94}]$	前後の移動量の平均

これによって，どの期間のユーザ行動を使うことによって，各推定モデルに対しどの程度の精度が得られるかの評価を可能とする．

5.2　推定モデル

抽出した特徴量を利用して $RQ1$〜$RQ3$ それぞれを検証するための推定モデルを生成し評価した．具体的には $RQ1$ に対しては，ユーザの行動データに基づく適合性推定の推定モデル $P_{S/E}$，$RQ2$ に対しては，不適合の原因を推定する推定モデル $P_{I/S}$，及び $RQ3$ に対しては，適合するドキュメントを推定する推定モデル P_{RD}，を取得したユーザの行動ログを利用して生成，評価した．

$P_{S/E}$ の推定モデルの構築及び評価には，適合したドキュメントを発見できた場合のユーザの行動ログと適合したドキュメントを発見できなかった場合のユーザの行動ログが必要である．前者は，クエリ q_2 が発話されず，なんらかの適合ドキュメントがユーザによって選択された場合の行動ログ U_S を，後者はクエリ q_2 が発話され再検索された際のユーザの行動ログである U_{IEE}，U_{VE} を利用した．

$P_{I/S}$ の推定モデルの構築及び評価には，適合したドキュメントを発見できず，なおかつその際の原因が音声エラーに起因する場合とその際の原因が検索意図との相違に起因するユーザの行動ログが必要である．前者は，クエリ q_2 が発話され，なおかつ提示したドキュメント集合が D_{IEE} の場合，後者はクエリ q_2 が発話され，なおかつ提示したドキュメント集合が D_{VE} の場合を利用した．

P_{RD} の推定モデルの構築及び評価には，適合したドキュメントを発見できた場合のユーザの行動ログが必要である．クエリ q_2 が発話されず，すなわち適合するドキュメントが発見された際のユーザの行動ログをそれぞれ適合するドキュメントの番号 (6 件) で分類したデータを利用した．

以上の行動データを 5.1 節で説明した手順に基づき学習データ及びデータセットを生成した．

推定モデルの学習モデルは，勾配ブースティングの一種であり，弱識別器に決定木を用いた Gradient Boosting Decision Trees を利用した．Gradient Boosting Decision Trees はブースティングを利用しており，与えた教師付きデータを用いて学習を行い，その学習結果を利用して逐次的に重みの調整を繰り返すことで複数の学習結果を求め，その結果を統合・組み合わせ，精度を向上させるという手法を取る．また，損失関数に決定木・回帰木など樹木モデルを採用しており，これによって，特徴量が連続変数・カテゴリ変数であっても対応が容易で，外れ値や欠損値にも強いという特徴がある．Grand Boosting Decition Tree におけるパラメータの設定は Tree-structured Parzen Estimator Approach [1] によって決定した．データセット及び学習データは 10-fold crossvalidation によって各データセットを分割した上で，推定モデル及び評価を実施した．

推定モデルの精度の評価尺度は，Accuracy を用いる．ここで，Accuracy は $P_{S/E}$ は，正解ラベルであるユーザの検索行動が検索意図に適合しているかどうか，$P_{I/S}$ は，正解ラベルである不適合の理由が検索意図に適合していないからか，それとも音声検索のエラーに起因しているか否か，P_{RD} は，6 種類のドキュメントからどのドキュメントを適合としたか，を正しく推定できた割合，と定義される．

推定モデル $P_{S/E}$，$P_{I/S}$，P_{RD} は，各特徴量単体で作成したモデル，及び特徴量 **EG**，**EM**，**FP**，**HP**，及び特徴量 **TT** を組みあせたモデル，及び特徴量 **EG**，**EM**，**FP**，**HP** を組み合わせたモデル，及び特徴量 **EG**，**EM**，**FP**，**HP**，**TT** を組み合わせたモデルを各区間毎，及び各区間を組み合わせたモデル毎に構築した．

推定モデルのベースラインは，$P_{S/E}$，$P_{I/S}$ は **Random**，**TT** の 2 つの手法を，また，P_{RD} は，**Random**，**TT**，**Select all 1** の 3 種類を用意した．**Random** はその性質から $P_{S/E}$，$P_{I/S}$ の Accuracy は 0.5，P_{RD} の Accuracy は 0.167 である．また，**TT** では特徴量として f_1 のみを利用し，すなわちユーザの行動ログから取得できる特徴量を利用する手法である．**Select all 1** は，検索結果からすべて一番目のドキュメントを選択する手法である．これは，ユーザは適合するドキュメントが検索結果に複数含まれる場合，上位の検索結果ほどユーザがクリックしやすい，という知見に基づく．同手法による Accuracy は 0.212 である．

(注4)：Accuracy における太字は実験条件下で最も有意な値を得られた値とす

表 2　推定モデル $P_{S/E}$ の分類性能 (注4)

	Session 1	Session 2	Session 3	average(1,2,3)	Session 1+2	Session 1+2+3
Random [BL1]	0.500	0.500	0.500	0.500	0.500	0.500
TT[BL2]	–	0.592	0.592	–	0.592	0.592
EG	0.535	0.555	0.782	0.624	0.520	0.747
EM	0.575	0.524	0.780	0.626	0.522	0.765
FP	0.573	0.629	**0.819**	**0.674**	**0.642**	**0.846**
HP	0.508	0.525	0.775	0.603	0.476	0.749
EG+EM+FP+HP	0.583	0.635	**0.815**	**0.678**	**0.627**	**0.846**
EG+TT	–	0.610	0.776	0.693	0.612	0.814
EM+TT	–	0.604	0.761	0.682	0.582	**0.837**
FP+TT	–	0.652	**0.813**	**0.732**	**0.656**	0.831
HP+TT	–	0.575	0.798	0.686	0.584	0.794
EG+EM+FP+HP+TT	–	0.658	**0.827**	**0.743**	**0.658**	**0.856**

表 3　推定モデル $P_{I/S}$ の分類性能 (注4)

	Session 1	Session 2	Session 3	average(1,2,3)	Session 1+2	Session 1+2+3
Random[BL1]	0.500	0.500	0.500	0.500	0.500	0.500
TT[BL2]	–	0.603	0.603	–	0.603	0.603
EG	0.620	**0.642**	0.603	**0.622**	**0.628**	0.590
EM	0.599	0.638	0.620	0.619	0.615	0.590
FP	0.563	0.576	0.600	0.580	0.563	0.559
HP	0.590	0.589	0.615	0.598	0.560	**0.602**
EG+EM+FP+HP	0.605	0.585	**0.623**	**0.604**	**0.614**	**0.609**
EG+TT	–	0.633	0.633	**0.633**	**0.645**	0.625
EM+TT	–	0.589	**0.642**	0.616	0.615	**0.645**
FP+TT	–	0.600	0.632	0.616	0.572	0.568
HP+TT	–	0.615	0.637	0.626	0.573	0.615
EG+EM+FP+HP+TT	–	0.571	**0.641**	**0.606**	**0.605**	**0.600**

表 4　推定モデル P_{RD} の分類性能 (注4)

	Session 1	Session 2	Session 3	average(1,2,3)	Session 1+2	Session 1+2+3
Random[BL1]	0.167	0.167	0.167	0.167	0.167	0.167
Select all 1[BL3]	0.223	0.223	0.223	0.223	0.223	0.223
TT[BL2]	–	0.234	0.234	–	0.234	0.234
EG	0.274	0.302	**0.345**	0.307	0.290	0.282
EM	0.242	0.274	0.341	0.286	0.298	0.290
FP	0.216	0.291	0.340	0.282	**0.345**	**0.324**
HP	0.326	0.278	**0.345**	**0.316**	0.266	0.306
EG+EM+FP+HP	0.204	0.270	**0.332**	**0.269**	**0.237**	**0.258**
EG+TT	–	0.294	0.222	0.258	0.298	0.286
EM+TT	–	0.274	0.234	0.254	0.270	0.274
FP+TT	–	**0.316**	0.254	**0.285**	**0.303**	**0.312**
HP+TT	–	0.266	0.246	0.256	0.286	0.310
EG+EM+FP+HP+TT	–	**0.299**	0.233	**0.266**	**0.245**	**0.254**

5.3 考　　察

推定モデルの評価を行った結果を表 2，3，4 に記す．まず $RQ1$ に対する考察を行う (表 2 参照).

EG，**EM**，**FP**，**HP** の中で最も効果が高かった特徴量は average(1,2,3) の結果によると **FP** であり，Accuracy は 0.674 である．**FP** は対象区間における顔の器官点の移動量の平均を取ったものであり，特に $[f_{71} - f_{91}]$ の特徴量が高い寄与をなしていることが確認されており，これは，検索行動中の口の動き，例えばユーザの独り言等の行動が反映された可能性がある．また，**FP** に次いで効果が高かった特徴量は average(1,2,3) の結果から分かる通り **EM** であり Accuracy は 0.626 である．同結果を検証するため，ユーザの顔画像の動画を視認し確認したところ，検索結果において適合する検索結果が提示された際に，比較的感情が笑顔になったり，適合しない検索結果が提示されなかった場合，うんざりしたりした表情が確認された．

次に特徴量を組み合わせた場合について考察する．

最も高い Accuracy を得られた項目は，特徴量として **EG+EM+FP+HP+TT** を利用しさらに全区間の特徴量を利用した場合であり，Accuracy は 0.856 である．BL2 の行動履歴を利用し推定した場合には，Accuracy の値が 0.592 であることから，我々の提案手法である，ユーザの行動履歴とその他の行動から得られる特徴量を組みあせて適合性を推定する，という提案手法の優位性が示せたことを意味する．また，本推定モデルで推定した適合性の推定結果は，検索結果のランキング学習で利用する場合，ユーザが再クエリを入力する前に検索結果の適合性が推定できることが望ましく，Session1+2 の期間でいかに良い推定結果が得られるかが重要となる．同観点からすると，**EG+EM+FP+HP+TT** における Accuracy の値が 0.627 という値が得られている．同値は BL2 の値より高い推定値が得られている．

また，今回のタスクはユーザに **TT** はユーザの操作の履歴を一部利用している．しかしながら，音声検索では，音声検索のみでタスクを実行することもあり，その際は当然操作履歴を利用することはできない．よって **TT** を利用せずに適合性を判定できることが望ましい．同観点から考えると，Session 1 + 2 で **TT** の Accuracy は 0.592 であったのに対し，**EG+EM+FP+HP** は 0.627 を得ることができた．同結果から，ユーザの操作履歴無しでも音声検索中のユーザの行動を得ることでより高精度に適合性を推定できることがわかる．

次に，$RQ2$ に対する考察を行う (表 3 参照).

EG，**EM**，**FP**，**HP** の中で最も効果が高かった特徴量は average(1,2,3) の結果によると **EG** である．特に Session1, Session2 における **EG** の値が高い値を示している．ユーザの顔画像の動画を視認して検討したところ，特にこれはクエリ q_1 の認識結果が間違っているとユーザが判断した途端，画面の凝視をやめて画面以外の場所を見る，等のユーザ行動が確認された．

次に特徴量を組み合わせた場合について考察する．

$P_{I/S}$ における推定結果の利用用途はクエリ修正を想定している．同理由からユーザが再発話をする前，すなわち Session1+2 の期間で適合性の推定を行えるか否かが最も重要である．その点，**EG+EM+FP+HP** の特徴量を利用した際に Accuracy が 0.64 と，BL1，BL2 ともに上回る結果が得られていることがわかる．

最後に，$RQ3$ に対する考察を行う (表 4 参照).

同推定結果は，P_{RD} と同様に検索結果のランキング学習をオンラインで行う用途を考えた場合，Session2 以前の特徴量を用いて適合性の推定ができることが望ましい．同観

る

点で考えると，Session 1+2 における **FP** を利用した場合に，Accuracy が 0.343，また，**FP+TT** を利用した場合 0.303 といずれも他の特徴量を利用した場合と比較して若干精度が高いことが確認された．これは検索結果が表示された画面を閲覧する際に，適合する検索結果等を注視する，等のユーザ行動が出た可能性が考えられる．

ただし同実験結果は BL を大幅に上回ることはできていない．これは学習データに利用したデータ数が比較的少ない事や特徴量の設計，ユーザの行動を記録する記録時間等の影響が出ている可能性がある．同課題を修正した上で再検証する必要があると思われる．

6. ま と め

本論文では，音声検索における適合性推定実現を目指し，音声検索時のユーザの無意識な行動を利用した適合性推定の実現性検討を行った．

その結果，音声検索中のユーザの無意識な行動を利用することによって，検索結果に対する適合性を推定することができることを示した．特に，音声検索中のユーザの顔の器官点の動き及び視線を観察することによって，より効率的に検索結果に対する適合性を推定できることを示した．また，検索結果に対する不適合の理由の推定については，ユーザの感情を観察することによって，より効率的に適合性を判定できることを示した．また，ユーザの無意識的な身体的行動を複合的に組み合わせて推定することによってより高い精度で適合性を判定できることを示した．

今後の課題は，同推定結果を利用した検索結果の精度向上方法の提案，及びクエリ修正手法の提案等が挙げられる．

文　献

[1] J. Bergstra. Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems 24*, pages 2546—-2554, 2011.

[2] C. Braunagel, W. Stolzmann, E. Kasneci, and W. Rosenstiel. Driver-activity recognition in the context of conditionally autonomous driving. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 1652–1657. IEEE, 2015.

[3] H. T. Dang, D. Kelly, and J. J. Lin. Overview of the trec 2007 question answering track. In *TREC*, volume 7, page 63, 2007.

[4] M. J. Eugster, T. Ruotsalo, M. M. Spapé, I. Kosunen, O. Barral, N. Ravaja, G. Jacucci, and S. Kaski. Predicting term-relevance from brain signals. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 425–434, New York, NY, USA, 2014. ACM.

[5] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 34–41, New York, NY, USA, 2010. ACM.

[6] A. Hassan Awadallah, R. Gurunath Kulkarni, U. Ozertem, and R. Jones. Characterizing and predicting voice query reformulation. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 543–552. ACM, 2015.

[7] S. Huffman. How voice search will change digital marketing — for the better. May 5th, 2016 from `https://moz.com/blog/how-fvoice-fsearch-fwill-fchange-fdigital-fmarketing-for-fthe-allowbreakbetter`.

[8] S. Huffman. Omg! mobile voice survey reveals teens love to talk. Retrieved October 14, 2014 from `https://googleblog.blogspot.jp/2014/10/omg-mobile-voice-survey-reveals-teens.html`.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 154–161, New York, NY, USA, 2005. ACM.

[10] M. P. Kato and K. Tanaka. To suggest, or not to suggest for queries with diverse intents: Optimizing search result presentation. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 133–142, New York, NY, USA, 2016. ACM.

[11] J. Kincaid. Google:25% of queries from android 2.0 devices use voice search. Retrieved July 18, 2014, from `http://techcrunch.com/2010/08/12/googles-hugo-barra-25-of-android-queries-are-voice-based/`.

[12] Y. Moshfeghi and J. M. Jose. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, pages 133–142, 2013.

[13] M. Nunez. All the new features for your mac. Retrieved June, 2016, from `http://gizmodo.com/macos-sierra-first-impressions-what-its-like-to-use-si- 1782023959`.

[14] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the ntcir-9 intent task. In *NTCIR*. Citeseer, 2011.

[15] K. Umemoto, T. Yamamoto, S. Nakamura, and K. Tanaka. Search intent estimation from user's eye movements for supporting information seeking. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 349–356, New York, NY, USA, 2012. ACM.

[16] K. Umemoto, T. Yamamoto, S. Nakamura, and K. Tanaka. Predicting query reformulation type from user behavior. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 894–901, New York, NY, USA, 2013. ACM.

[17] H. Wang, Y. Song, M.-W. Chang, X. He, A. Hassan, and R. W. White. Modeling action-level satisfaction for search task satisfaction prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 123–132, New York, NY, USA, 2014. ACM.

[18] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 297–306. ACM, 2006.