社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

頭部モーションセンサと音声を用いた対話インタフェースの検討

會田　卓也† 　西本　卓也†† 　大川　茂樹† 　嵯峨山茂樹††

† 千葉工業大学
†† 東京大学

あらまし　我々は，視覚によらずに情報機器を操作する手段として，音声認識や音声合成が広く受け入れられるための総合的なヒューマンインタフェースのあり方を検討しており，特に，頭部モーションセンサにより認識したユーザの頭部動作を音声入出力と組み合わせる手法を検討している．今回は，頭部の角度および角速度を状態遷移モデルによって認識しつつ，誤認識を防ぐ配慮を行った頭部認識モジュールを実装し，評価を行ったので報告する．
キーワード　頭部モーションセンサ, 音声入力, 音声出力, 音声認識

A Voice Interface System with the Head Motion Sensor

Takuya AIDA†, Takuya NISHIMOTO††, Shigeki OKAWA†, and Shigeki SAGAYAMA††

† Chiba Institute of Technology
†† The University of Tokyo

Abstract　We are investigating a human machine interface with speech recognition and speech synthesis as a means to operate information devices without using visual display. This report describes the technique of combining voice input and output with actions of the user's head which are recognized by the motion sensor attached to the head. In this method, angles and angular velocities of the head are recognized by the state-transition model. As the experimental result, the subjects were able to input two style of gestures correctly with the system after the short practice. The sound effects which give feedbacks to the users and prevent incorrect recognitions were also effective.
Key words　Head Motion Sensor, Voice Input, Voice Output, Speech Recognition

1. はじめに

音声入力によってシステムに指示を行ない，システムからの音声出力のみによってフィードバックや情報を得る，といったインタラクションは，特に車載音声インタフェースや視覚障害者の支援技術など，ユーザが視覚的なフィードバックを得にくい場合に有効である．このような場面において，要素技術としての音声認識がより広く受け入れられるためには，音声認識を有効に使うための，総合的なヒューマンインタフェースのあり方を考える必要がある．

入力手段としての音声認識には，手や視覚を拘束されないという利点がある．また，大語彙かつ自然な発話を正しく認識できれば，効率的で自然な入力インタフェースとなることが期待される．しかし多くの場合には，音声による入力に加えて，発話の開始をシステムに知らせたり，認識結果として得られる複数の候補から適切な項目を選択したり，入力された内容をキャンセルしたり，誤認識の訂正などを行なう，といった操作が必要となる．音声認識の結果がたとえ完璧であっても，入力途中でユーザの気が変わった，といった場合もありうる．したがって，このような補助的な入力手段は，将来，音声認識性能が十分に向上した場合であっても重要であろう．

従来，音声入力を用いたシステムにおいては，スイッチやキーボード，ポインティングデバイスなど他の入力手段を併用することが多い．このようなシステムは，操作のタイミングに慣れることが難しかったり，音声入力とデバイス操作の組合せが繁雑になったりする．また，特に視覚を用いない音声のみによる操作環境やウェアラブルコンピュータなどの環境では利用しにくくなる．

本研究では，音声対話に伴って人間が頭部を動かす動作に着目し，3 次元モーションセンサを用いて頭部動作を測定し，これを音声入力と組み合わせて利用する，という入力インタフェースについて検討する．ただし，必ずしも自然な人間の動作の認識を対象とするのではなく，習熟が容易で，頑健で確実な入力手段を検討する．また，効果音によって動作認識システムの内部状態をユーザに提示することの有効性についても合わせて検討する．

我々が用いる 3 次元モーションセンサは圧電振動ジャイロ，加速度センサ，地磁気センサなどを利用しており，小型かつ軽

量である．これを頭部に装着することで，音声対話に伴う人間の自然な頭部動作を測定できる．モーションセンサは音声や画像などと比較して周囲の環境の影響を受けにくいため，得られる値からの動作の認識は比較的頑健に行なえると考えられる．

本報告では，ユーザの音声入力に対するシステムからの応答を聞きながら頭を縦や横に振るなどして，肯定や否定などの意志表示を行なう入力インタフェースの提案を行う．

2. 関 連 研 究

音声認識と他の入力手段を併用するシステムとしては，音声認識とセンサを併用する MIT の Put-that-there システム [1] をはじめ，タッチパネルの併用 [2], [3]，マウスやキーボードの併用 [4] など，さまざまな提案がなされている．これらは，音声認識という入力手段の利点を生かしつつ，音声入力では困難な空間的な位置や座標の入力，誤認識時の訂正など，音声入力の弱点を補うものである．また，モーションセンサをヒューマン・マシン・インタフェースに利用するシステムとして，Toss-it [5] がある．これは，相手の PDA にボールをトスするかのように自分の PDA を振ることで，直感的な情報の移動を行うものである．

3. 頭部運動を用いた対話の提案

本研究では，人間が音声で対話を行うときに頷きや首をかしげる動作などを認識することで，操作方法を学ぶ労力を軽減しつつ，複数モダリティを併用する入力システムを実現したい．

人間同士の対話において，肯定・否定の頭部動作は，「はい」「いいえ」などの発話と同時に生じることが多い．また，相手が自分の声を聞き取れない場合などには頭部動作のみでも情報が伝わる．対話における確認のやりとりをマルチモーダル化することにより，ユーザにとって信頼できる入力システム，つまり，効率性や確実性の高いシステムを実現できる可能性がある．

インタフェースシステムに動作認識を用いる場合は，

（1） 人間の自然な動作をできるだけ高精度に認識する．

（2） システムが確実に認識できるような動作を対象とし，ユーザがそのような動作を確実に行えるように練習をする．

という 2 つのアプローチが考えられる．ユーザに強いる負担をできるだけ減らす，という立場からは，前者が理想的である．しかし，入力そのものが自然な行為であっても，誤認識によって使い勝手が損なわれる場合には，ユーザの負担は大きくなる．これはすでに音声認識の応用において生じている問題である．第 4 章の実験では以下のことを確認する．

- 人間同士の対話における自然な頭部動作は，同じ意図の動作においても個人差やバリエーションが大きく，頑健なパターン認識は容易ではない．

本研究は音声認識を補完する入力手段を目指すために，後者のアプローチを選ぶ．その際，ユーザの負担を軽減するために，以下の配慮が重要ではないかという仮説を立てた．

（1） 人間の自然な動作にできるだけ近い動作を用いることで，ユーザの学習が容易になる．

（2） 用いる動作を，ユーザが意識しやすい複数の状態に分割することで，ユーザの学習が容易になる．（例：「頭を下げる」「しばらく待つ」「頭を上げる」）

図 1 帽子に取り付けた 3D モーションセンサ

（3） 特に重要な状態遷移が起きた場合には効果音を提示することで，ユーザの学習が容易になり，確実な入力が可能になる．

第 5 章の実験ではこれらの仮説を検証する．

4. 頭部運動データの予備的検討

4.1 センサーハット

本研究で使用する NEC Tokin 製 3D モーションセンサ MDP-A3U9S の仕様を以下に示す．

- ロール角 (X 軸) 検出範囲：± 180 deg
- ピッチ角 (Y 軸) 検出範囲：± 90 deg
- ヨー角 (Z 軸) 検出範囲：± 180 deg
- インタフェース：USB 1.1
- 外形寸法：20 mm × 20 mm × 15 mm
- 重量：6 g
- 対応 OS：Windows 98/Me/2000/XP

本研究では，モーションセンサを図 1 のように帽子の頭頂部に付け（以下，「センサーハット」と呼ぶ），29.97 サンプル／秒でデータを取得した．このセンサは，ロール角（X 軸），ピッチ角（Y 軸），ヨー角（Z 軸）を取得することができる．

否定および肯定の頭部動作を順番に行った場合の角度とその差分・二次差分の値の例を図 2〜4 に示す．図 2（角度値）においては，否定の動作（首を横に振る）でヨー角が正および負に 1 回ずつ変化する．また，肯定の動作（首を縦に振る）でピッチ角が負に 1 回変化する．得られる値は連続的であり，センサの出力値にはノイズがほとんどない．図 3（角度の差分）においては，否定の動作でヨー角が「正→負→正」のパターンで変化する．また，肯定の動作でピッチ角が「負→正」のパターンで変化する．

他のセンサの利用についても検討したが，光学式モーションセンサの場合は角度値の誤差が大きいなどの問題があった．本研究で用いたモーションセンサは比較的安価で，小型・軽量であり，十分な精度で角度値を得ることができる．

4.2 自然な頭部運動データの収集

人間が相手に肯定や否定の意思を自然に伝えている状況の頭

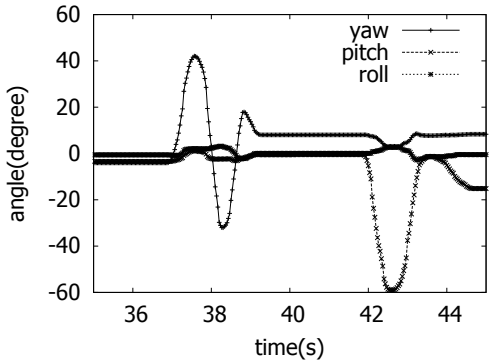

図 2　3D モーションセンサの出力例（角度）

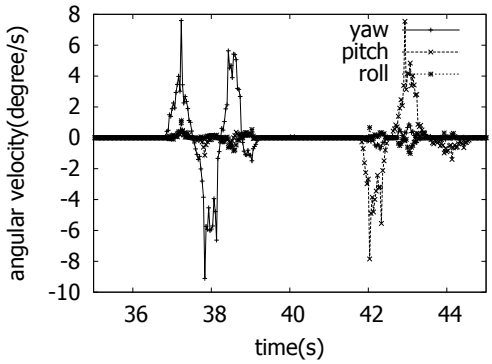

図 3　3D モーションセンサの出力例（角速度）

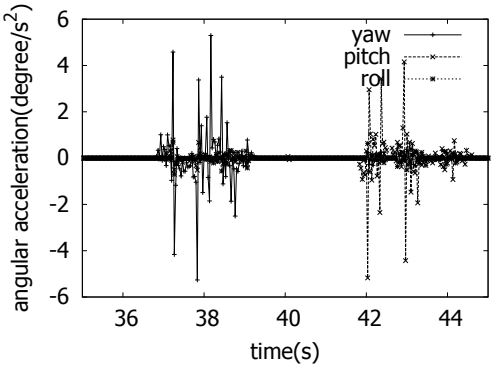

図 4　3D モーションセンサの出力例（角加速度）

部運動を取得した．

カーナビにおける目的地の地名やランドマーク名など（Point of Interest, 以下，POI）を選択するタスクを設定した．タスクでは，実際にカーナビの音声入力の性能評価用に設計した 152 個の POI を分類した地名リスト [7] を用いた．

このタスクでは実際に人間が読み取れるように出題者が頭部動作を行い，回答者も実際にその動作を読み取っているため，人間同士の自然な頭部動作を取得できる．また，複数の試行を行っても，慣れの効果などによって難易度が変わったり，頭部動作のスタイルが変化しにくいと考えられる．

被験者は 20 歳から 24 歳の理工系学生 6 組 12 人（男性 7 人，女性 5 人）である．ただし，センサを装着した被験者（出題者）は，男性 4 人，女性 2 人（日本人 4 人，中国人 2 人）である．国籍の違いは特に実験に支障がないと判断し，実験はすべて日本語で行った．

表 1　HMM(3 状態 1 混合) による認識率

	サンプル数	正解数	認識率 (%)
Yes	59	46	77.97
No	156	144	92.31
Idle	213	210	98.59
計	428	400	93.46

表 2　正解数と認識結果数 (3 状態 1 混合)

	結果 Yes	結果 No	結果 Idle
正解 Yes	60	1	15
正解 No	2	155	8
正解 Idle	0	1	214

6 組の被験者から取得したデータは合計約 30 分，各組のセッションは平均約 5 分である．得られた動作サンプルの数を以下に示す．

- 肯定 (Yes)：59 個
- 否定 (No)：156 個
- その他の状態 (Idle)：213 個

4.3　HMM による頭部運動データの認識

音声や映像と同期づけながら目視でラベリングを行い，上記の動作サンプルの切出しを行う．これら 3 カテゴリのパターンを隠れマルコフモデル (HMM) により認識する実験を行った．ただし，パラメータは 3 軸の角度のフレーム間差分値（3 次元ベクトル）である．学習と評価にはすべてのデータを行ったため，closed test での評価である．状態数は 2〜5，各状態のガウス分布の混合数は 1,2,4 のいずれかとし，それぞれの条件で比較した結果，3 状態 1 混合モデルにおいて特に良好な結果を得た．認識率を表 1 に，各動作サンプルがどのように認識されたかを表 2 に示す．

表 2 の結果より，Yes を Idle と誤認識する場合がやや多かったが，No および Idle は比較的正しく認識できた．No の動作は Yes に誤認識されることがほとんどない．一方 Yes はモデルの精緻さが不十分である可能性がある．

本研究の目標としては，音声認識をシステム全体として信頼性の高い入力手段にすることである．わずかな確率でも誤認識が生じると，システム全体をユーザが信頼できなくなる．この実験の結果から，自然な頭部運動の認識を対象として，動作の切り出しを確実に行い誤認識を防ぐことは難しいと考えられる．

5.　頭部運動データの認識

5.1　認識手法の検討

先行研究では，モーションセンサの値から動作を認識するにあたって「誤認識を防ぐアルゴリズム」を提案しているものに Toss-it がある．Toss-it は基本的に速度の積分値の比較を行い頑健性を保障している．

我々は Toss-it と同じ手法を検討したが，手首ではなく頭を動かすという前提の違いから，動作の速度や精度が違うために同じ手法は適切ではないと判断した．

そこで，頭を縦または横に傾けて戻すという一連の動作をい

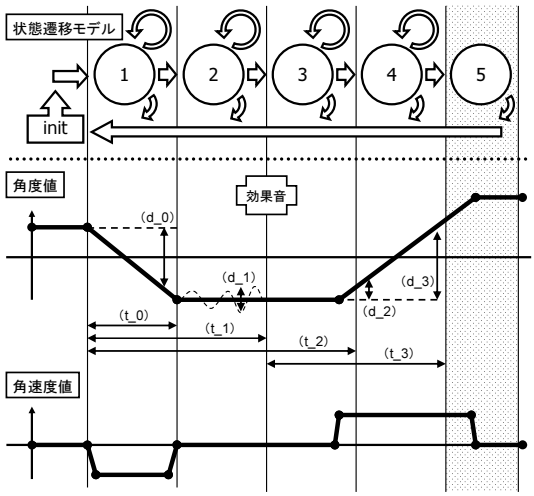

図 5　頭部の角度および角速度の状態遷移モデル

くつかの状態に分割し，それぞれの状態遷移に時間や角度の閾値を設定する，という状態遷移モデルを用いて，Toss-it での速度の積分値との比較と同等な頑健性を目指すことにした．その際,「意図した入力を確実に行う」「意図しなかった入力を確実に回避する」の両者についてユーザがコツをつかみやすくすることを重視する．

5.2　状態遷移モデルによる認識手法

図 5 にこのアルゴリズムで用いた状態遷移モデルの概要を示す．このモデルでは，頭部の角度および角速度によって状態が遷移する．図中の角度値，角速度値における矢印は閾値を表しており，縦の矢印は角度値，横の矢印は時間を意味している．なお，簡略化のため角度値，角速度値は等速度運動に単純化してある．

遷移の条件を表 3 に表す．現在の角度値を「D」と表し，時間値を「T」と表す．なお，状態が遷移する際の最終的な角度値を「D_状態番号」として表し，角速度値を「V_状態番号」として表す．例えば，状態が 1 から 2 に遷移する際の最終的な角度値は「D_1」とする．また，遷移時の条件として設定した角度値における閾値は「d_閾値番号」として表し，時間値における閾値は「t_閾値番号」として表す．初期化時の状態番号は 0 とする．

図 5 の状態遷移モデルについて，状態 5 は受理を意味している（以下,「Accept」と呼ぶ）．また，状態が 2 から 3 に移る際に（表 3 におけるイベント 5），状態 4 に移ることができる合図である効果音（以下,「効果音 A」と呼ぶ）が鳴る．全ての状態は初期化を経て状態 1 に遷移することができる（以下,「Reject」と呼ぶ）．

このアルゴリズムを用いて，頭の頷き（以下,「Positive」と呼ぶ）とかしげ（以下,「Doubtful」と呼ぶ）のジェスチャーの認識システムを実装した．開発にはモーションセンサの SDK が提供する API を使用し，Microsoft Visual C++ 6.0 を用いてコマンドラインのプログラムを作成した．動作環境は Microsoft Windows XP である．

Positive の認識には Pitch の角度値，角速度値を用い，Doubtful には Roll の角度値，角速度値を用いた．利用者は頷きをし，効果音 A が聞こえたときに頭を上げることで入力することができる．もし，途中で入力の意思を変えたときや誤入力が起きたときは，少しの間頭を動かさないことで入力をキャンセルすることができる．

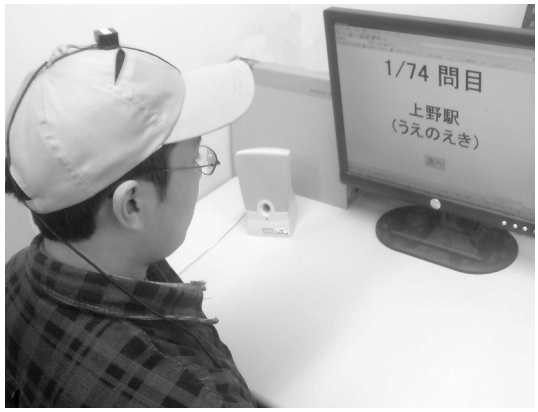

図 6　実験の様子

なお，状態 3 および 4 から動作をキャンセルした場合（表 3 におけるイベント 10,11,12）には Reject を示す効果音（効果音 B）が鳴る．また，状態 5 に到達した場合（表 3 におけるイベント 13）には Accept を示す効果音（効果音 C1/C2，それぞれ Positive と Doubtful に対応）が鳴る．これらにより，Accept や Reject が確実に行われたことを，ユーザは効果音によって知ることができる．

5.3　状態遷移モデルを用いた実験

肯定や否定の意思をコンピュータに入力するタスクを設定した．入力手法は Positive と Doubtful の二種類とする．POI を照合するタスクとして，以下のような課題を行わせる．

（1） 被験者はセンサーハットをかぶり，PC の前に座る．

（2） 被験者は 1 分間程で自分にあった入力のタイミング（閾値）を選ぶ．

（3） モニターに地名が表示され，スピーカーから地名を表す合成音声が流れる．

（4） 被験者は表示された地名と聞えた音が同じかどうかを，頭の動きだけで PC に入力する．同じであった場合は Positive を，異なっていた場合は Doubtful を入力する．

（5） 3, 4 を全ての問題が終わるまで繰り返す．

実験の様子を図 6 に示す．実験に用いた地名は，事前に合成音声として作成したものを用いる．また，被験者にはウェブブラウザ上で入力を行ってもらう．被験者の入力を JavaScript を用いて検知することで，被験者は自分自身で問題を進めることができる．

問題は 74 問あり，答えの半分が○で半分が×である．被験者は 21 歳から 25 歳までの今までに実験に参加したことのない理工系学生 8 人（男性 6 人，女性 2 人）である．

5.4　実験結果

8 人の被験者から取得したデータは合計約 35 分，各被験者のセッションは平均約 4 分である．各被験者は 1 回の実験で 74 回の入力を行った．得られた動作サンプルがどのように認識さ

表 3　状態遷移の条件

イベント番号	状態の変化	条件
1	1 → 初期化 → 1	T - T_0 が t_0 より大きいとき
2	1 → 2	イベント 1 でなく，D - D_0 の絶対値が d_0 より大きく，T - T_0 が t_0 より小さいとき
3	1 → 1	イベント 1,2 でなく，T - T_0 が t_0 より小さいとき
4	1 → 初期化 → 1	イベント 1,2,3 でないとき
5	2 → 3	T - T_1 が t_1 より大きいとき
6	2 → 初期化 → 1	イベント 5 でなく，D - D_1 の絶対値が d_1 より大きいとき
7	2 → 2	イベント 5,6 でないとき
8	3 → 4	D - D_2 の絶対値が d_2 より大きいとき
9	3 → 3	イベント 8 でなく，T - T_2 が t_2 より小さいとき
10	3 → 初期化 → 1	イベント 8,9 でないとき
11	4 → 初期化 → 1	T - T_0 が t_2 より大きいとき
12	4 → 初期化 → 1	イベント 11 でなく，T - T_2 が t_3 より小さいとき
13	4 → 5	イベント 11,12 でなく，D - D_3 が d_3 より大きいとき
14	4 → 4	イベント 11,12,13 でない場合
15	5 → 初期化 → 1	無条件

表 4　状態遷移モデルにおける結果数

	結果 Positive	結果 Doubtful
正解 Positive	296	0
正解 Doubtful	0	296

表 5　状態遷移モデルにおける Reject の発生回数

	Positive	Doubtful
効果音 A 前 Reject	49	160
効果音 A 後 Reject	48	190

れたかを表 4 に示す．また，Reject がどのような状態で発生したかを表 5 に示す．

表 4 の結果より Positive，Doubtful 共に意図しない誤入力は発生することはなく，頑健に動作した．

表 5 の結果より Doubtful は Positive と比べ Reject の回数が多くあった．被験者の中には，入力の Reject を自分自身の意思で行えない被験者がいたが，意図しない入力の際に，頭の動きが無意識に止まっていたために，正しく Reject することができていた．

被験者ごとの Reject の回数を図 7 に示す．Reject が発生する回数は被験者によって大幅な差があり，少ない人で 1 回，多い人で 205 回であった．全ての被験者は初心者であるため，Reject が多い人はセンサーハットのコツと最初に触ったときのイメージが離れていると考えられる．

練習によるなれの効果を知るため，Reject の回数が多い被験者 2 人に対して約 5 分の練習を行ってもらい再実験を行った．その結果を図 8 に示す．練習では我々が被験者に口頭でコツの教示のみを行った．

本結果では，Reject の回数が 8%, 20% と非常に減った．このことより，僅かな練習で頑健に入力できるようになることがわかる．

本実験では実験中にタスクと無関係な動作を被験者にあまりさせていない．今後は，例えば入力システムを利用しないで人間同士で自由に会話を行っているような場合に，意図しない動作の受理を防ぐことが容易であるかどうか，といった検討が必要である．

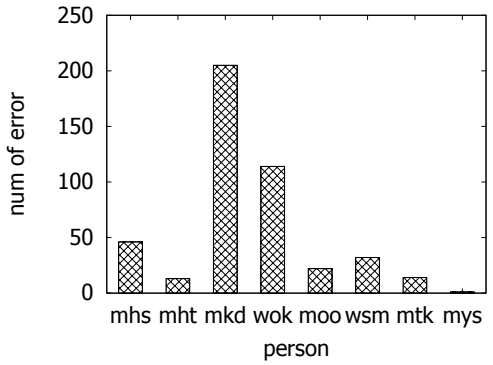

図 7　74 回の入力に対して発生した Reject の回数

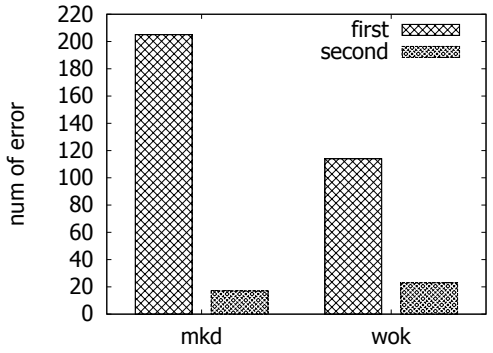

図 8　練習を行った際の Reject の回数の推移

5.5　フィードバックの必要性

効果音 A の有無によって，認識にどのような影響を与えるかを実験した．被験者は Reject の少ない 2 名を選び，状態遷移モデルでの状態が 2 から 3 に遷移する際の効果音 A を鳴らさずに再実験を行った．再実験を行うにあたって，約 5 分間被験者に自由に練習を行わせた．その実験結果を図 9 に示す．また，被験者 mys の Reject の回数の変化を図 10 に示す．図 10

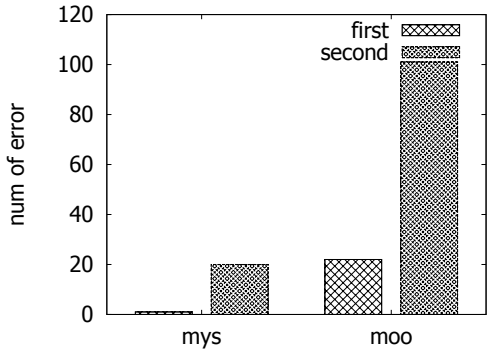

図 9　効果音 A の有無による Reject 数の推移

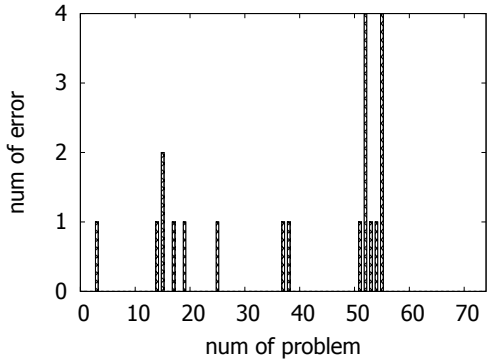

図 10　被験者 mys の Reject の回数の変化

の縦軸は一回の入力あたりの Reject の回数を表している．

効果音 A を鳴らさないことによって Reject の回数が 20.0 倍, 4.6 倍と増加した．また，図 10 より Reject が発生する場所に偏りがあり，この傾向は被験者 moo にも見ることが出来た．

Reject が偏ってしまった原因は，

（1） 被験者は効果音 A を聞くことが出来ないため，図 5 における状態 2 から 状態 3 へ遷移したタイミングを知ることが出来ない．

（2） 頭を戻すタイミングを被験者自らの勘で判断し入力を行わなければならず，問題を繰り返すうちにタイミングがずれてしまい Reject が発生してしまう．

（3） 効果音 A が無いために被験者は Reject の発生した原因を知ることが出来ず，再び同じ原因で Reject を発生させてしまう．

ということだと予測できる．被験者の勘によりタイミングを判断することは不可能ではなかったため，なれの効果により効果音 A を除去しても思い通りに入力できるかもしれないが，現段階では内部の動作を知るための効果音 A は，確実性を保障するために必要であるといえる．

現在の状態をユーザが知ること，つまりユーザがフィードバックを得ることは学習を容易にさせ，認識率を向上させるために重要であるといえる．

6. ま と め

本報告では，頭部運動を 3 次元モーションセンサにより取得し，これを音声入力と組み合わせることにより，自然で頑健な入力を実現する音声対話インタフェースを提案した．また，頭部運動をモーションセンサにより取得する実験を行い，得られたデータを遷移モデルで認識する実験を行い，本手法の有用性を確認した．

本手法を視覚障害者支援や車載音声インタフェースとして用いるためには，頑健性を保障したまま利用しやすいインタフェースにしていく必要がある．今後は，ユーザが容易に利用できるように，ユーザごとに閾値などの設定を数回の練習のみで最適化できるシステムの構築を目指す．

本報告で述べた動作認識モジュールは，音声対話システム開発ツール Galatea Toolkit [8] と組み合わせて有効性の評価を行う予定である．特にキーボードやマウスが使用できない状況で，音声入力と頭部モーションセンサ入力を併用し，音声出力によって情報を提示するシステムにおいて，入力の効率性やユーザの負荷軽減といった観点から，総合的なヒューマン・インタフェースとしての有効性を検討していきたい．

謝辞

本研究の一部は，文部科学省特定領域研究「情報福祉の基礎」視覚障害班（計画研究班キ, Kiki 班）および科研費若手研究 (B) 「テレマティックス音声対話における安全性と快適性の評価に関する研究」の成果である．特に，モーションセンサに関してご教示をいただいた慶應義塾大学・安村通晃教授と塚田浩二氏の両氏に感謝します．また，実験に協力してくださった千葉工業大学・大川研究室の皆様に感謝します．

文　献

[1] Bolt, R. A. : Put-that-there : Voice and gesture at the graphics interface, ACM Computer Graphics, Vol. 14, No. 3, pp. 262-270 (1980).

[2] 竹林洋一 :音声自由対話システム TOSBURG II －ユーザ中心のマルチモーダルインタフェースの実現に向けて－, 電子情報通信学会論文誌, D-II, Vol. J77-D-II, No. 8, pp.1417-1428 (1994).

[3] 神尾, 松浦, 正井, 新田 : マルチモーダル対話システム Multiks-Dial, 電子情報通信学会論文誌, D-II, Vol. J77-D-II, No. 8, pp. 1429-1437 (1994).

[4] 西本, 志田, 小林, 白井 : マルチモーダル入力環境下における音声の協調的利用 –音声作図システム S-tgif の設計と評価, 電子情報通信学会論文誌, D-II, Vol. J79-D-II, No. 12, pp. 2176-2183 (1996).

[5] Miyahara, K., Inoue, H., Tsunesada, Y., Sugimoto, M. : Intuitive Manipulation Techniques for Projected Displays of Mobile Devices, In Proceedings of ACM CHI2005 Extended Abstract, Portland, Oregon, pp.1881-1884 (2005).

[6] 會田, 西本, 中沢, 大川, 嵯峨山 : 頭部モーションセンサと音声を用いた対話インタフェースの提案, ヒューマンインタフェースシンポジウム 2005 講演論文集, 2531, pp.601-604, (2005).

[7] 西本, 西村, 赤堀, 石川, 磯谷, 伊藤, 大淵, 金澤, 國枝, 外山, 新田: 音声認識応用に関する学会試行標準, 情報処理学会研究報告, 2005-SLP-55, pp. 47-52, (2005).

[8] http://hil.t.u-tokyo.ac.jp/~galatea/