論文 / 著書情報
Article / Book Information

論題(和文)	音声とペンの同時入力における個人差への適応化
Title(English)	
著者(和文)	渡邉 康司, 篠田 浩一, 古井貞熙
Authors(English)	Yasushi Watanabe, Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会2008年春季講演論文集, Vol. , No. 2-4-11, p. 55-58
Citation(English)	, Vol. , No. 2-4-11, p. 55-58
発行日 / Pub. date	2008, 3

音声とペンの同時入力における個人差への適応化*

◎渡邉康司, 篠田浩一, 古井貞熙 (東工大)

1 はじめに

PDAや携帯電話などのモバイル端末が普及し，メールなどの文章を入力しやすいインタフェースが求められている．それらの文章を入力する方法としては，音声，手書き文字，テンキーなどがある．音声は理想環境下では90%以上の認識性能を持ち，入力速度はキー入力よりも速い．しかし，音声認識の性能は周囲雑音の影響を受けやすく，周囲雑音が大きいモバイル環境では著しく性能が劣化する．一方，PDAなどでしばしば利用される手書き文字などのペン入力は，一般に認識性能は高いものの入力速度が遅い．テンキーも同様に入力速度は遅く，長い文章を入力するには不向きである．このように音声入力とペン入力は一長一短がある．音声とペン入力から得られる情報は独立と考えられるため，それらを組み合わせることにより，音声のみの入力に比べ耐雑音性に優れ，ペン入力のみに比べ入力が速いというインタフェースが期待できる．

我々は，大語彙連続音声認識とペン入力を組み合わせた，音声とペンの同時入力を用いたインタフェースを開発した [1]．これは，話しながらペン入力を行うというインタフェースである．同時入力方法は，そこで用いられていた方法以外にも考えられるため，本稿では，モバイル端末で使われるキーボードを用いた入力を提案する．また，それらの同時入力インタフェースにおいては，ユーザーは音声とペンの入力をできるだけ同期して入力を行う．しかし，実際には，二つのモード間でずれが生じる．それのずれは，話者ごとに異なっている．さらに，音響的特徴についても話者ごとに異なっている．それらの個人差に頑健な認識を行うために，適応を行う．

2 同時入力インタフェース

入力方法は，先行研究で特に効果の高かった「手書き文字による文節先頭文字の入力」と「文節先頭文字の行の選択」および，本稿で新たに提案する「キーボードを用いた入力」の3種類を用いる．ここでは，それらの入力方法について述べる．

いずれの入力でも，ユーザーは音声を発声しながら，文節の先頭においてペン入力を行う．ここで，音声は連続音声である．ただし，全ての文節の先頭で入力を行う必要はなく，ユーザーが入力可能なタイミングでペン入力を行う．その際ユーザーは，文節ごとの音声入力開始とペン入力開始のタイミングが，できるだけ一致するようにする．

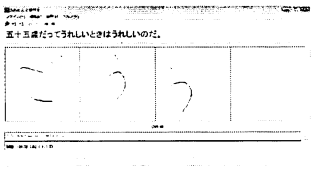


Fig. 1 先頭文字の入力

Fig. 2 文字テーブル

1. 手書き文字による文節先頭文字の入力 (Fig. 1)
   文節先頭文字の読みを手書き文字により平仮名で入力する．入力インタフェースには4個の枠があり，左から一枠に一文字ずつ入力する．一番右まで書いた後は，また左から入力する．

2. 文字テーブルによる文節先頭文字の選択 (Fig. 2)
   手書き文字を書かずに，文字テーブルを用いてペンタップにより文節先頭文字の「行」を選択する．文字テーブルとは，50音表の「行」ごとにボタンを配置したものである．例えば，「お」という文字を入力したいときは，あ行の領域をタップする．

3. キーボードの配列を用いたインタフェース (Fig. 3)
   一般的なキーボードの配列であるQWERTY配列を用い，文節先頭文字の読みをローマ字表記に変換し，その最初の1文字を入力する．例えば，「五十」(ゴジュー) と入力するときは，「ゴ」に対応するローマ字表記「GO」の最初の1文字「G」を入力する (Fig. 4)．端末には，小型PCである「SONY VAIO type U」を使用する．4.5型ワイドのタッチパネル液晶を持ち，重さは500グラムである．キーボードはソフトウェアキーボードを使用する．単語の先頭になり得ない「Q」と「X」は，押されても入力とは見なさない．入力にはペンではなく指を用いてもよい．実際の収録では，主に指が使われた．本稿では，指による入力を含めて「ペン入力」と呼ぶこととする．

---

*Model adaptation for semi-synchronous speech and pen input
By Yasushi Watanabe, Koichi Shinoda and Sadaoki Furui (Tokyo Institute of Technology)

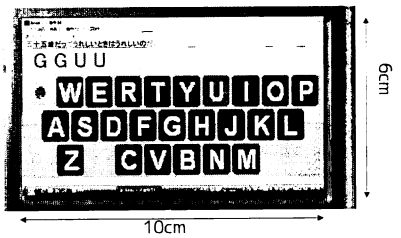


Fig. 3　キーボードを用いたインタフェース

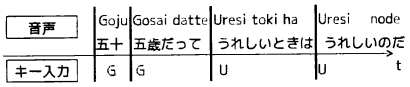


Fig. 4　キーボードの配列を用いたキー入力の例

3　マルチモーダル認識

本研究では，先行研究 [1] で用いられていたマルチモーダル認識アルゴリズムを用いる．そこでは，2パス処理により認識を行う．第1パスではオンラインで音声認識尤度とペン入力認識尤度を組み合わせて認識を行い，単語グラフを生成する．第2パスでは，音声とペン入力のずれが正規分布に従うと仮定し，この正規分布を用いて重み付けをしたペン入力認識尤度を，単語グラフに反映させる．以下，第1パスのみ簡単に説明する．

入力においては，文節の発声開始時刻とペンの入力開始時刻をできるだけ一致させるようユーザーに指示しているが，実際には文節の発声開始時刻とペン入力開始時刻の間にはずれが生じる．ペン入力認識尤度を，意図した音声認識の単語仮説に反映するには，そのずれを考慮する必要がある．そこで，ペン入力開始時刻から対応する単語の開始時刻を引いたものを入力のずれ μ と定義する．そして，μ だけペン入力開始時刻を前に補正する (Fig. 5)．ここで，μ は負の値も取り得る．

次に，音声認識の単語仮説のなかで，その区間がペン入力開始時刻に含まれるものを列挙する．それらの単語仮説に対し，式 (1) を用いて，ペン入力認識尤度を反映させる．

$$L = L_s + \alpha L_p \ . \tag{1}$$

ここで，L_s は単語仮説の尤度，L_p はペン入力認識尤度，α は重み係数である．このとき，音声認識のプロセスはペン入力認識の結果が得られるまで，ペン入力開始時刻で停止していることとする．アルゴリズムのフローチャートを Fig. 6 に示す．以上の処理により，ペン入力認識尤度が高い文字を先頭文字とする単語仮説が，ビーム幅内に残る可能性が高くなるので，解候補を効率的に絞り込むことが可能になる．

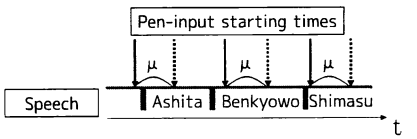


Fig. 5　ペン入力開始時刻の補正の例 (破線の矢印が補正前のペン入力，実線の矢印が補正後のペン入力を表す．)

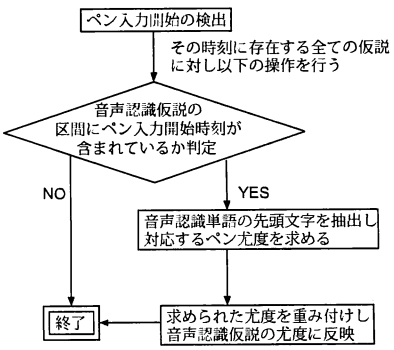


Fig. 6　ペン入力認識尤度を音声認識の単語仮説に反映させるアルゴリズム

4　個人差への適応化

4.1　入力タイミングのずれ

音声とペン入力のタイミングのずれは話者ごとに異なっている．そのため，3章で述べた μ の値は，話者ごとに調節する必要がある．音声とペン入力のタイミングのずれを求めるためには，ペン入力がどの単語に対し入力されたかを決定する必要がある．先行研究 [1] では，人手により対応づけを行い，ペン入力のずれ μ の値はその対応付けの情報を用いてテストデータから求めていた．本稿では，その対応付けを自動で行い，適応により μ を求める．以下に教師あり適応の手順を示す．

今，一つの文入力において，C 個のペン入力と M 個の単語が入力されたとする．ペン入力の各々を $c_n (n = 1, \cdots, C)$，それらの入力開始時刻を t_n とする．まず，適応データに対し，正解ファイルを与えて強制切り出しを行うことにより，各単語仮説 $w_m (m = 1, \cdots, M)$ の発声開始時刻 t_m を求める．次に，ペン入力 $c_n (n = 1, \cdots, C)$ に対応する単語を以下の手順で決定する．

Step. 1 文節先頭になり得ない助詞や助動詞は候補から除く．

Step. 2 $|t_n - t_m| < I$ を満たす単語仮説のなかで，ペン入力 c_n に対応する文字を先頭文字としてもつものを列挙する．ここで，I は正の定数である．その条件を満たす単語仮説が存在しなかった場合は，Step.4 に進む．

Step. 3 Step.2 で列挙された単語仮説のうち，$|t_n - t_m|$ が最小となる単語仮説とペン入力 c_n を対応

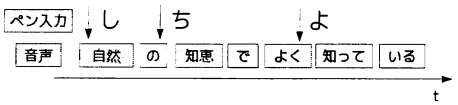


Fig. 7　音声とペン入力の対応づけの例 (各ペン入力は灰色で表された単語と対応づけられる．)

付けて，処理を終了する．

Step. 4 $|t_n - t_m|$ が最小となる単語仮説 w_m を c_n と対応付ける．

ここで，I は，音声とペン入力のずれについての閾値であり，Step.2においては，I 以上ずれた入力は考えないものとしている．また，「キーボードの入力」では，例えば「G」という入力なら，「が」「ぎ」「ぐ」「げ」「ご」がペン入力に対応する文字となる．「手書き文字の入力」では，手書き文字の認識結果の上位 N 文字をペン入力に対応する文字とする．上位 N 文字に正解が含まれる確率がほぼ100%となるように N を設定する．

ペン入力の開始時刻を t，そのペン入力と対応づけられた単語の開始時刻を t' とすると，$\Delta t = t - t'$ がペン入力のずれとなる．ユーザーの適応データにおける Δt の平均値を，マルチモーダル認識時に μ として用いる．

教師無し適応を行う場合は，正解ファイルを与える代わりに，音声認識を行う．認識結果の一位の結果から，各単語の発声開始時刻を求める．その結果をもとに，教師あり適応と同様の方法で音声とペン入力の対応付けを行い，Δt を求め，ユーザーごとの平均値をマルチモーダル認識時に μ として用いる．

4.2　音響モデル

音響的な特徴はユーザーごとに異なっている．また，マルチモーダルインタフェースを使用するときと通常の発声においても音響的な特徴が異なっていると考えられる．そこで，音声認識精度の向上のために音響モデルの適応を行う．MLLR [2] を使って HMM を適応し，マルチモーダル認識では，適応された HMM を用いる．

5　評価実験

5.1　収録条件

2章で述べた3つインタフェースについて，録音室での日本人20名 (男性19名，女性1名) の収録を行った．入力形態ごとに日本音響学会の音素バランス文からなる研究用連続音声データベース (ASJ-PB) と新聞記事読み上げ音声コーパス (ASJ-JNAS) から無作為に抽出した50文を入力した．このうち，最初に入力した15文を適応データ，残りの35文を評価データとして用いる．また，被験者は収録前に各インタフェースを10分程度使用し，ある程度慣れた状態で収録を行った．また，タイミングが明らかにずれたり，明らかに誤った入力をした場合は，録り直しを行った．

5.2　実験条件

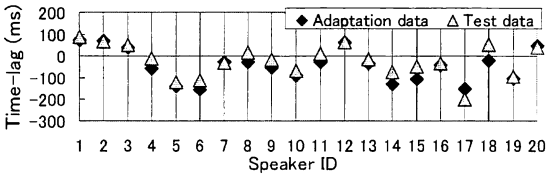
音響モデルは連続音声認識コンソーシアム2000年度版ソフトウェア [4] に含まれている triphone HMM を用いた．単語辞書は毎日新聞の1995年から2001年までの記事データから，読みの存在しない記号等を取り除き，出現頻度上位60,000単語から作成した．言語モデルは単語辞書と同様のデータを用いた3-gramである．

手書き文字モデルは，「オンライン手書き文字パターンデータベース」[5] における，被験者10名の平仮名計43,800文字を用いて学習された連続型 HMM を用いた．認識単位をストロークとしており，ストローク単位数は25，各ストロークあたりの状態数はペンダウン状態が3(自己遷移ありスキップなし)，ペンアップ状態が1(自己遷移なし) である．また，各状態あたりの混合成分数は1である．手書き文字辞書は，平仮名計71字について，25方向のストローク列で表記したものを作成した．辞書作成の際に，筆者により筆跡が大きく異なるものはパターンを複数用意した．その結果手書き文字は82文字となった．

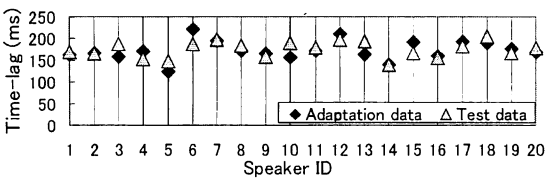
今回はシミュレーション実験として，通常版の Julius で認識して求めておいたペン入力認識尤度や時間情報を音声認識時に読み込み，認識を行った．重み係数 α は事後的に最適な値を使用した．音響モデルの適応には，教師あり MLLR を用いた．話者ごとのペン入力のずれの適応については，教師あり適応により求めた μ_s を用いた場合，教師無し適応により求めた μ_u を用いた場合をそれぞれ行った．教師あり適応については，話者ごとに適応データを用い，教師なし適応については，話者ごとにすべての評価データを用いた．4章で述べた，ペン入力のずれの適応で使われるパラメータについては，$I = 300$ms，$N = 10$ とした．これは，先行研究で使用していたデータベースにおいて，音声とペン入力のずれはほぼ300ms以下であったこと，手書き文字認識において，第10位までに正解が含まれる確率が97%であったことからその値に設定した．また，ペン入力のずれについての不特定話者認識のために，ある話者について，他のすべての話者の適応データから求めたずれの平均 μ_0 を用いて認識するということを，すべての話者について繰り返した場合の実験も行った．

5.3　実験結果

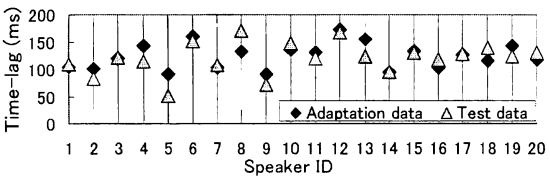
Fig. 8に話者ごとの適応データと評価データのぞれぞれについて，音声とペン入力の時間のずれの平均値



(a) 手書き文字



(b) 文字選択



(c) キーボード

Fig. 8　音声とペン入力の時間のずれ

を示した．ずれの平均値は話者により異なっている．適応データと評価データでのずれがほぼ同じ値となった話者がいる一方で，そうでない話者もいる．

Fig. 9 に，音声認識およびマルチモーダル認識の単語正解精度を示した．それぞれのインタフェースについて，共通の音響モデルを用いた場合と音響モデルを適応した場合を示している．「文字選択」と「キーボードの入力」では，教師あり適応よりも教師なし適応が高い単語正解精度となった．適応データは最初に入力をした 15 文であり，評価データは，その後に入力した 35 文である．入力を行う間に話者のペン入力のずれの傾向が徐々に変わり，評価データを用いた教師なし適応の方が，評価データのペン入力のずれの傾向をより正確に表した結果と考えられる．μ_0 を用いた場合とペン入力のずれ μ について適応を行った場合を比較すると，適応の効果がある場合と効果がない場合がある．教師なし適応については，手書き文字において音響モデルを適応した場合を除き，単語正解精度が改善しており，最大で 0.5 ポイントの単語正解精度の改善が得られた．

音響モデルの適応を行うことにより，音声認識の単語正解精度が 2.5–2.6 ポイント改善した．音響モデルの適応とマルチモーダル認識を組み合わせることで，最大で，4.5 ポイントの改善が得られた．キーボードの入力が単語正解精度の改善が最も大きいという結果になった．音声とペン入力のマルチモーダル認識に音響モデルの適応を組み合わせることは効果的である．

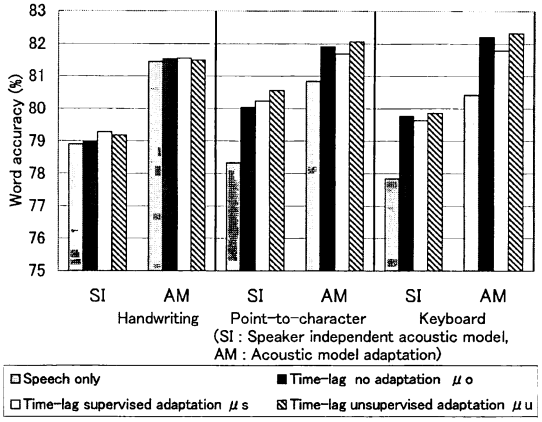


Fig. 9　認識結果

6　おわりに

本研究では，音声とペンの同時入力における個人差に対し頑健な認識を行うため，音声とペン入力のずれの適応および音響モデルの適応を行った．3 種類の入力方法について，モバイル端末を用いて収録したデータで評価実験を行い，適応の効果があることを示した．話者ごとに適応を行い，マルチモーダル認識をすることで，最大で 4.5 ポイントの単語正解精度の向上が得られた．今後の課題としては，マルチモーダル認識における重み係数 α の適応が挙げられる．

謝辞　本研究は，文科省科学研究費補助金 (基盤 B, 課題番号 15300054) の助成を受けた．オンライン手書き文字データベースを提供して頂いた東京農工大の中川研究室に感謝する．

参考文献

[1] Y. Watanabe, et al., "Semi-Synchronous Speech and Pen Input", *Proc. ICASSP 2007*, Hawaii, U.S.A., SPE-L5.1 (IV-409), 2007.

[2] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continu ous density hidden Markov models," Computer Speech and Language, Vol. 9, No. 2, pp. 171–185, 1995.

[3] 李, "大語彙連続音声認識エンジン julius の開発と進展", 情処研報 SLP-59, 127-132, 2005.

[4] 河原 他, "連続音声認識コンソーシアム 2000 年度版ソフトウエアの概要と評価", 情処研報, SLP-38-6, 37-42, 2001.

[5] 中川 他, "文章形式字体制限なしオンライン手書き文字パターンの収集と利用", 信学技報 PRU95-115, 43-48, 1995.