社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005—NL—167（5）
2005—SLP— 56（5）
2005／5／26

音声と手書き文字の同時入力インターフェース

中川竜太　　小林唯　　小林隆二　　篠田浩一　　古井貞熙

東京工業大学
〒 152-8552 東京都目黒区大岡山 2-12-1
Email: {rtag, furui}@furui.cs.titech.ac.jp, {yui, kobayasi, shinoda}@ks.cs.titech.ac.jp

「書きながら話す」「話しながら書く」という音声と手書き文字の同時入力インタフェースをマルチモーダル認識により実現する．音声のみの入力に比べ耐雑音性に優れ，手書き文字のみの入力に比べ入力速度が大きいという特色をもつ．複数モードの統合方法として，従来マルチモーダル認識に用いられてきた事前統合や事後統合ではなく，オンラインで統合を行いながらサーチを行う中間的な統合方式を採用する．これにより，文・文章など比較的長い入力への対応が可能となり，音声と手書き文字の同期のずれに対し頑健性が高くなることが期待される．本稿では，音声認識の結果として出力された単語グラフにおける尤度に手書き文字認識の尤度を反映させる 2 パス処理を用いてその可能性を検証した．被験者ごとに同期のずれの補正を行うことにより，音声のみの認識性能から単語正解精度で 2.6 ポイント改善した．

Simultaneous Input Interface of Speech and Handwritten Characters

Ryuta Nakagawa, Yui Kobayashi, Ryuji Kobayashi,
Koichi Shinoda, and Sadaoki Furui

Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan

We propose an interface using simultaneous inputs of speech and handwriten characters. This interface is more robust against noise than the speech interface, and its input speed is faster than the inteface with handwritten characters. For integrating the two modes, we employ a method which fusions the tentative recognition results from them on-line during the recognition process, which is different from the convensional integration methods performed in feature level or in semantic level. The proposed method is applicable to long inputs such as sentences and expected to be robust against the large asynchronousity of the two inputs. In this paper, the proposed method was preliminarily evaluated by using a two-pass process in which a word graph generated by speech recognition in the first pass is utilized for the integration process of the two modes in the second pass. The proposed method improved the recognition accuracy by 2.6 point over the method only with speech recognition.

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．音声は理想環境下では 90%以上の認識性能をもち，入力速度はキー入力よりも速い．しかし，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認識は，認識性能は高いものの入力速度が遅い．このように音声入力と手書き文字入力はどちらも一長一短である．そこで，ここではメール文などの自由文入力を対象とした，音声と手書き文字との同時認識を行う「書きながら話す」「話しながら書く」入力インタフェースの実現を目指す．このインタフェースは音声のみの入力に比べ耐雑音性に優れ，手書き文字のみの入力に比べ入力速度が大きいという特色を持つ．また，本研究は，手書き文字認識における予測インタフェースにおいて，音声を用いてその予測精度を高める研究と位置づけることも可能である．

本研究はマルチモーダル認識研究の一つと位置づけられる．複数モードを統合する手法は事前統合 (もしくは特徴レベル統合) と事後統合 (もしくは意味レベル統合) に大別される．事前統合の代表的な研究としては，口唇の画像認識 (リップリーディング) を音声認識とともに用いるバイモーダル認識がある [1]．そこでは，フレーム長程度の同期のずれに対応する Factorial HMM などの認識アルゴリズムが研究されている．しかしこのような事前統合方式では，ある程度以上の時間間隔の同期のずれに対応することは難しい．一方，事後統合の例としては，音声とジェスチャー (簡単な図形の入力) を組み合わせて認識する研究がある [2]．この手法は，両入力の同期を前提としているため簡単なコマンドの認識に利用範囲が限られており，自由文を認識対象とする本研究では，直接用いることができない．そこで，本研究では，オンラインで複数モードの統合を行いながらサーチを行う中間的な統合方式を採用する．これにより，文・文章など比較的長い入力への対応が可能となり，音声と手書き文字の同期のずれに対し頑健性が高くなることが期待される．

本論文では，その第一段階として，まず音声認識を用いて単語グラフを出力し，さらに手書き文字認識も併用してグラフ探索を行う 2 パス処理の枠組みで評価を行った．手書き文字の入力は音声に比べると非常に遅く，同一の情報を同時に入力することはできない．入力インタフェースとして適切な手書き文字を選別する必要がある．例えば漢字での入力は，平仮名に比べて画数が多いため，入力が遅いという欠点がある．また，文節の区切り情報のような入力もあるが，情報量が少ないため大幅な性能改善が望めない．そこで，ここでは文節先頭の読みを平仮名で入力するインタフェースを採用した．

2　マルチモーダル認識

本研究では音声，手書き文字ともに HMM によりモデル化する．音声は音素を，手書き文字はストローク (画) を認識単位とする．音声と手書き文字の同期のずれはその分布を確率モデルで表現する．音声入力のサーチ途中で非同期に入力される手書き文字入力による尤度を反映させる方式を目指す．ここでは予備検討として 2 パス処理により認識を行う．つまり，まず音声認識の結果を単語グラフとして出力する．そして手書き文字入力が対応する単語に，手書き文字尤度を重み付けで加える．最後に全ての手書き文字入力の尤度を反映させた単語グラフをリスコアリングして，最も尤度の高い認識結果を得る．ここで問題となるのは，どのように単語グラフ上の単語に手書き文字を対応付けるかである．

今，一つの文入力において，入力された手書き文字が C 個あったとし，各々を $c_n(n = 1, \ldots, C)$ する．また，手書き文字 c_n の入力の開始時刻を t_{c_n} とする．さらに，音声認識で出力された単語グラフの各々のアークを $w_m(m = 1, .., M)$ とする．このとき，各々のパスにおいて，手書き文字 c_n の入力開始時刻 t_{c_n} に最も近いノード q_l を開始時刻に持つアーク w_m に，手書き文字 c_n の尤度を重み付けて加えることとする．

$$L'(w_m) = L(w_m) + \alpha H(c_n) \tag{1}$$

ここで，$L_(w_m)$ は音声認識による単語 w_m の尤度，$H(c_n)$ は手書き文字入力 c_n に対する尤度であり，α は音声認識と手書き文字認識との間の重み係数である．なお，この規則を単純に適用すると，1 つの手書き文字の尤度が複数のパスで加えられる現象が起き，特定のパスの尤度が不当に低くなる．これを防ぐために，一度手書き文字 c_n の尤度を反映させたアーク w_n にはフラグを立て，それ以上同じ手書き文字 c_n の尤度を反映させないようにする．例えば，図 1 は,「今日の 12 時集合です」と発声しながら文節の読みの先頭平仮名を「き (今日の)」「じ (12 時)」「し (集合です)」と手書き文字で入力したときの例である．この例では，手書き文字 c_2 の尤度は,「途中」,「中」,「12 時」に対応する単語アークに加えられる．以上の手法をここでは MTHD1 とする．

音声と手書き文字の入力開始時刻は常に異なっている．また，その異なり方は被験者に依存すると考えられる．すなわち，ある被験者は常に音声の入力開始が先であり，また，別の被験者は，手書き文字を先に書き始める，という現象が起きる．この入力時刻のずれを考慮することで，より効果的に手書き文字尤度を反映させることができると期待される．つまり，音声と手書き文字の同期のずれ $\delta = t_{c_n} - t_{q_l}$ をある分布関数 $p(\delta)$ に従う確率変数

とし，

$$L'(w_m) = L(w_m) + \alpha H(c_n) p(\Delta t) \qquad (2)$$

とする．ここで $p(\Delta)$ としては，正規分布を用い，そのパラメータは予めその被験者のデータから推定しておくこととする．以上の方法をここではMTHD2とする．

3　収録データ

先行研究 [3] では，音声と手書き文字の同時入力を収集したデータベースが存在しなかったため，音声と手書き文字それぞれ別々に収録されたデータを用いた．そして，手書き文字は全ての文節の先頭文字を入力し，対応する文節の発声開始時刻と完全に一致すると仮定したシミュレーション実験を行った．本論文では，より実際の利用に近づけるため，音声と手書き文字を同時に入力するインタフェースを試作し，音声と手書き文字の同時入力データを収録した．

3.1　収録インタフェース

データ収録には，WindowsXP を搭載した標準的なスペックのデスクトップパソコンを用いた．手書き文字入力のために，17 インチで横 1280 ドット，縦 1024 ドットの表示能力を持つ WACOM 社製タブレットモニタをこれに接続した．

収録用ソフトウェアは，[7] で配布されている Windows 用 JulisGUI 0.9.3 に，手書き文字入力部分を加えて作成した．手書き文字入力部は，120 ピクセル四方の枠が横に 10 個並べられており，それぞれ平仮名 1 文字を受け付ける．なお，画面上での枠のサイズは 3.1cm 四方である．被験者は，次節で述べる条件に基づき，発声開始ボタンを押してから左より任意の文節先頭文字を入力していく (図 2)．

インタフェースは，発声開始ボタンが押されると開始時刻を記録し，再度ボタンが押されるまでの間，音声と手書き文字入力を受け付ける．手書き文字は，枠内にプロットされるとその開始時刻を保存し，右隣の枠のプロットが開始されるまでの間，プロットされた場所の二次元座標，ペンの状態 (ペンアップ，ペンダウン) を保存する．プロットのサンプリング間隔は 20ms である．

3.2　収録条件

日本人男性 10 名について，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 96 文章，計 960 文の音声と手書き文字の同時入力データを収録した．収録は計算機や空調設備のある研究室内で行った．なお，収録前に以下の指示を被験者に与えた．

1. 文節などで区切らずに自然に発声すること
2. 任意の文節の読みの先頭ひらがなを手書き入力すること
3. 手書き文字は楷書でなくても構わないが，続け字にならないこと
4. 文節の発声開始時刻と手書き文字の入力開始時刻をできるだけ一致させること．発声終了時刻と手書き文字入力終了時刻は一致しなくても構わない

3.3　同時入力データの分析

音声と手書き文字の入力における同期のずれについて，収録データのうち，各被験者共通の 43 文 (形態素数 632) の分析を行った．音声の強制切り出しで得られた文節発声開始時刻と対応する手書き文字入力開始時刻との間のずれの平均と分散を表 1 に示す．常に音声よりも手書き文字入力が早くなる被験者や遅くなる被験者，ばらつきの大きい被験者がいることがわかる．前章で説明した同期のずれを考慮する手法 MTHD2 では，この値をパラメータとして用いる．なお, 同一の内容を発声してもらったにもかかわらず，入力された手書き文字の数は，被験者間で最大 2 倍以上の差がある．

4　評価実験

4.1　音声認識

IPA の「日本語ディクテーション基本ソフトウェア 1999 年度版」に収録されているモデルを用いた．音響モデルとしては，2,000 状態 16 混合性別非依存 triphone HMM を，言語モデルは毎日新聞 75ヵ月分の 2-gram モデルを，単語辞書は毎日新聞 45ヵ月

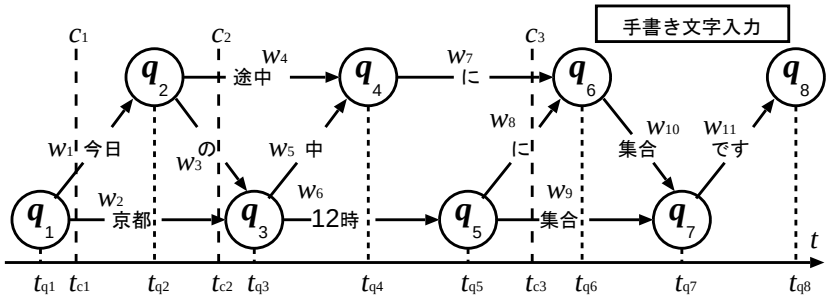

図 1: 単語グラフと手書き文字入力の例

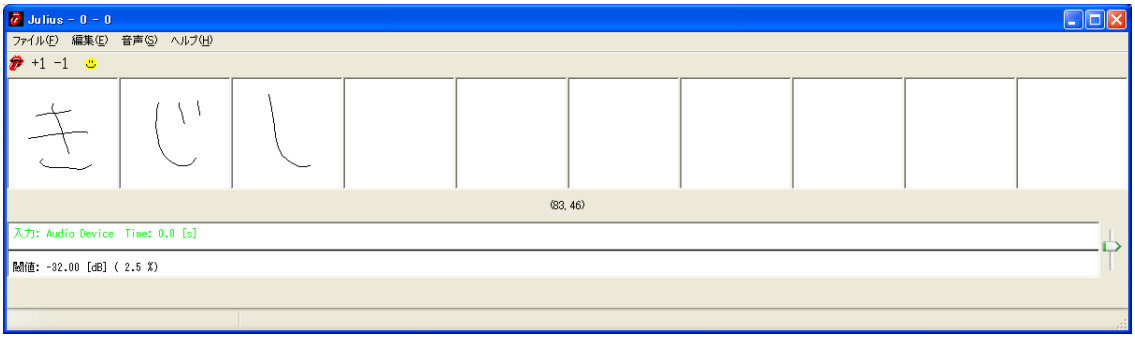

図 2: 収録例 「今日の 12 時に集合です.」

分の出現頻度上位 20,000 語をそれぞれ用いた．ただし，読みが存在しない単語 (句読点，記号など) に対しては，手書き文字尤度を反映させることができないため，これらの単語を予め辞書と言語モデルから削除した．また，この処理により，単語間にショートポーズがある場合に問題が生じるため，残りのすべての単語の先頭に，ショートポーズを表すスキップありの 1 状態 HMM を追加した．音響特徴量は 12 次元 MFCC とその微分，パワーの微分の計 25 次元を用いた．音声認識デコーダは HTK3.2 で，言語重みは 13.5，挿入ペナルティは 1.0 で固定した．

4.2 手書き文字認識

本研究の手書き文字認識手法として，嵯峨山らのストローク HMM 手法をを用いた [4],[5]．この手法では，認識単位をストローク (画) としており，崩れ字の認識，未学習字の認識，辞書登録による筆順違いへの対応などが可能となっている．特徴量は，ペン入力の x, y 方向の微分成分とペンの状態 (ペンアップ，ペンダウン) の 3 次元とした．ストローク HMM は，3 状態 1 混合とし，濁音，半濁音を含む平仮名 82 文字を [6] の被験者 10 人の平仮名計 43,800 文字のデータを用いて学習した．平仮名の平均ストローク数は 7.6，最長ストローク数は 25 である．

4.3 実験結果

収録したデータの中から各被験者共通の 10 文 (形態素数 209)，計 100 文を評価セットとして無作為に取り出した．3.3 節の分析で用いたデータとは異なる．評価セットにおける手書き文字認識の正解率 (1-best) と 10-best までの累積正解率 (10-best) を表 2 に示す．結果は，1-best でも正解率 87.4%と高く，10-best までの累積正解率では 96.6%となった．このことから，手書き文字認識は，話しながら書くという条件の下でも十分な性能が得られており，それを併用することにより，雑音環境下での音

表 1: 被験者ごとの音声と手書き文字の入力開始時刻のずれ (フレーム単位)

被験者 ID	0	1	2	3	4	5	6	7	8	9
入力文字数	123	152	101	231	162	105	96	133	129	138
平均	9.3	-7.8	4.4	-15.4	4.3	-1.7	5.1	-1.8	0.2	-1.5
分散	296.9	1004.8	206.2	342.4	257.2	206.0	61.8	354.6	298.3	230.2

声認識性能が向上する可能性があるといえる．

次に，評価セットを音声のみで認識したときの単語正解精度 (1-best) と 100-best までの累積単語正解精度 (100-best) を表 3 に示す．全体として 60.8%の単語正解精度が得られた．100-best までの累積単語正解精度は，74.3%であることから，2 パス処理での認識性能は最大で 14%程度改善する余地がある．

音声認識と手書き文字認識の 2 パス処理によるマルチモーダル認識の結果を，図 3 に示す．このとき，手書き文字尤度重み α は，被験者ごとに最も認識性能が高くなるものを採用した．全被験者の平均は，ベースラインである音声のみの結果 (SPCH) と比べ，手書き文字認識を加えた MTHD1 で 1.9%の単語正解精度が改善した．これにより，音声と手書き文字の同時入力によるインタフェースが，性能改善に寄与することが確認された．また，被験者ごとの音声と手書き文字の入力開始時刻のずれを考慮した MTHD2 では SPCH に比べ，2.6%改善した．なお，強制切り出しにより文節の音声開始時刻を求め，手書き文字の入力開始時刻と一致させて認識した結果 (ALGN) を調べたところ，MTHD2 よりもさらに 0.4%改善した．これらの結果は，音声と手書き文字の入力開始時刻のずれは被験者ごとに考慮する必要があることを示唆する．

次に，手書き文字尤度の重み係数 α を全被験者で共通としたときの結果を図 4 に示す．MTHD1 では，音声のみの結果を上回ることはなく，MTHD2 では $\alpha = 0.01$ で 0.4%の改善に留まった．これにより，重み係数 α は被験者ごとに最適化する必要があることがわかる．

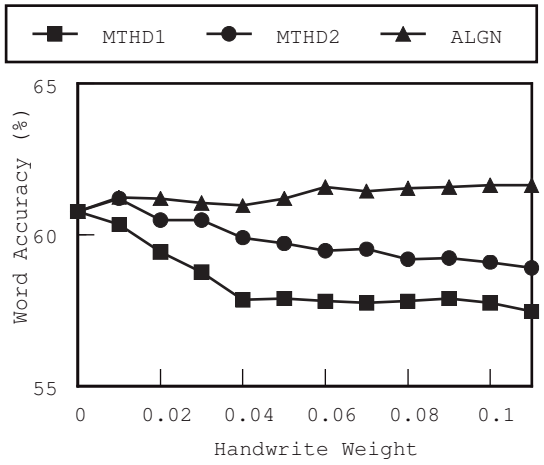

図 4: 手書き文字尤度重みと単語正解精度

5 まとめ

入力速度が大きく異なる音声と手書き文字の同時入力インタフェースを提案し，音声認識結果に手書き文字認識の尤度を反映する 2 パス処理でマルチモーダル認識を行った．音声のみの結果と比較し，認識性能の向上を確認した．また，被験者ごとの音声と手書き文字の入力開始時刻のずれに対処し，またモード間の重み係数を最適化することでより頑健な認識が行えることを示した．

今後，より使いやすいインターフェースの実現とさらなる性能向上のため，他の手書き文字入力の形態を検討する必要がある．例えば，文節の読みの先頭平仮名以外の情報，誤認識しやすい音韻や発声区間の始端，終端などがその候補である．同期入力のずれの分布の自動推定も，本稿の結果から重要であり，話者適応，筆者適応とともに組み入れる必要がある．さらに，雑音環境下での評価も行いたい．

また，本稿では同時認識アルゴリズムとして 2 パス処理を用いたが，今後，音声認識と手書き文字認識の探索アルゴリズムを統合することで，高速かつ高精度な認識手法を構築する．

表 2: 手書き文字認識の結果 (累積単語正解精度)

被験者 ID	0	1	2	3	4	5	6	7	8	9	all
1-best	86.1	86.4	88.2	76.7	89.1	92.9	97.0	92.3	92.1	85.4	87.4
2-best	91.7	88.6	94.1	82.2	89.1	92.9	100.0	97.4	94.7	87.8	90.8
...						...					...
10-best	94.4	95.5	100.0	93.2	95.6	96.4	100.0	100.0	100.0	95.1	96.6
手書き文字数	36	44	34	73	46	28	33	39	38	41	412

表 3: 音声のみによる認識結果 (累積単語正解精度)

被験者 ID	0	1	2	3	4	5	6	7	8	9	all
1-best	56.9	59.8	51.7	72.3	65.6	61.7	55.5	46.9	66.5	70.8	60.8
100-best	72.7	71.8	63.6	79.4	79.9	75.6	70.3	67.9	82.3	79.4	74.3

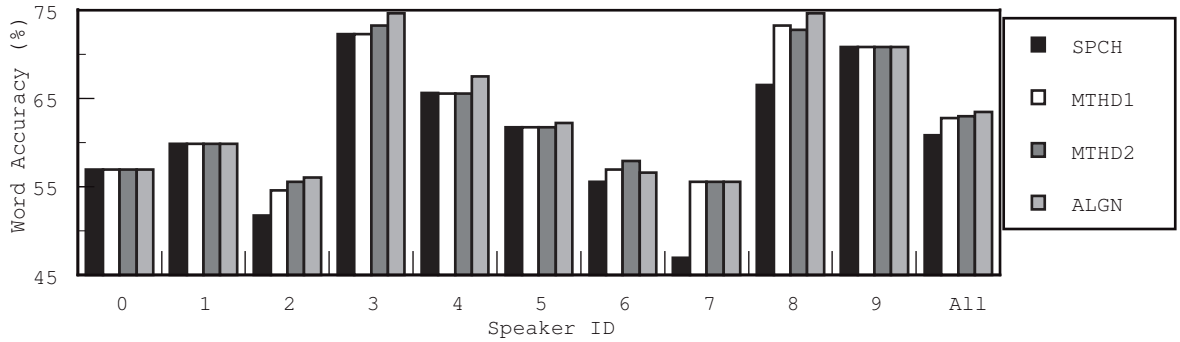

図 3: 音声と手書き文字同時認識の結果

謝辞

本研究は文部科学省科学研究費補助金基盤研究 (B)No.15300054 によるものである．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深く感謝する．

参考文献

[1] Satoshi Nakamura, “Statistical Multimodal Integration for Audio-Visual Speech Processing,” *IEEE Trans. Neural Networks*, vol.13, no.4, pp.854–866(2002).

[2] L. Wu, S. L. Oviatt, and P. R. Cohen, “Multimodal Integration-A Statistical View,” *IEEE Trans. Multimedia*, vol.1, no.4, pp.334–341(1999).

[3] 市屋 剛, 中川 竜太, 篠田 浩一, 古井 貞熙, “手書き文字の準同期入力を併用した音声認識手法の予備検討，” 電子情報通信学会 2004 年総合大会, D-14-007 (2004-3)

[4] 嵯峨山 茂樹, 中井 満, 下平 博, “ストローク HMM によるオンライン手書き文字認識方式,” 電子情報通信学会技術報告, PRMU2000–35, pp.1–8(2000-06).

[5] 中井 満, 嵯峨山 茂樹, 秋良 直人, 小場 久雄, 下平 博, “ストローク HMM によるオンライン手書き文字認識の性能評価,” 電子情報通信学会技術報告, PRMU2000–36, pp.9–16(2000-06).

[6] 中川 正樹, 東山 孝生, 山中 由紀子, 澤田 伸一, レー・バン・トゥー, 秋山 勝彦: “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,” 電子情報通信学会技術報告, PRU95–115, pp.43–48(1995-09).

[7] http://www.sp.m.is.nagoya-u.ac.jp/people/banno/julius_gui.html