論文 / 著書情報
Article / Book Information

論題(和文)	音声と手書き文字の同時入力によるインターフェースの検討
Title(English)	
著者(和文)	中川竜太, 小林唯, 小林隆二, 篠田浩一, 古井貞熙
Authors(English)	Ryuta Nakagawa, Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会2005年秋季講演論文集, Vol. , No. 1-7-11, pp. 13-14
Citation(English)	, Vol. , No. 1-7-11, pp. 13-14
発行日 / Pub. date	2005, 9

音声と手書き文字の同時入力によるインタフェースの検討*

◎中川竜太, 小林唯, 小林隆二, 篠田浩一, 古井貞熙 (東工大)

1 はじめに

「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検討する. 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. 入力速度の異なるこれら二つをマルチモーダル認識するため, 従来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頑健に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.

ここで, 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入力方法を選ぶ必要がある. 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.

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の時刻に対応する単語に, 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.

今, ある手書き文字 C の入力開始時刻が T であり, 単語グラフの i 番目の単語アークの音声の入力開始時刻 S_i と入力終了時刻 S_i が $S_i \leq T < E_i$ を満たすとする. この単語アークには単語 W_i が対応しており, その言語尤度は $L(W_i)$ であるとする (Fig. 1). そのとき, 単語アークの言語尤度を以下のように書き換える.

$$L'(W_i) = L(W_i) + \alpha H(X_i|C) \qquad (1)$$

ただし, α は手書き文字尤度重み, X_i は単語 W_i の先頭平仮名の読み, $H(X_i|C)$ は手書き文字 C が入力されたときの平仮名 X_i の対数尤度である. 単語グラフ中の全ての単語アークに対し, 入力された全ての手書き文字の対数尤度を反映させ, 得られた単語グラフを再探索し, 最尤の認識結果を得る. この手法をMTD1と呼ぶ.

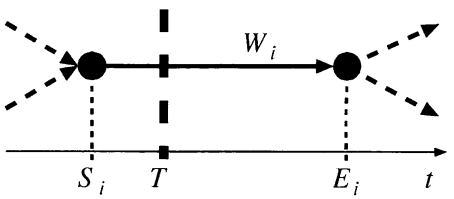

Fig. 1 A word graph and a handwritten character input.

一般に, 音声と手書き文字を同時に入力する場合, 両者の入力開始時刻を完全に一致させることは難しく, 前後にずれる. そこで, こ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. ここではまず, 手書き文字の入力開始時刻 T を予め推定した音声と手書き文字のずれの時間 $\overline{T}$ で $T' = T - \overline{T}$ として補正する. そして, $S_i \leq T' < E_i$ を満たす i 番目の単語アークの入力開始時刻 S_i に対し, 音声と手書き文字の同期のずれ $\delta = S_i - T'$ をある分布関数 $p(\delta)$ に従う確率変数として, 言語尤度 $L(W_i)$ を

$$L'(W_i) = L(W_i) + \alpha\{H(X_i|C) + \log p(\delta)\} \qquad (2)$$

と書き換える. ここで $p(\delta)$ としては, 正規分布を用い, そのパラメータは $\overline{T}$ とともに予め学習データから推定しておくこととする. この方法をMTD2とする.

3 評価実験

3.1 収録データ

音声と手書き文字の同時入力データを収録するインタフェースを作成し, 研究室内で日本人男性10名の収録を行った. 入力した文章は, ASJ-PB と ASJ-JNAS から無作為に抽出した各被験者共通の96文とした. このうち各被験者43文 (632形態素), 計430文を学習セットとし, MTD2で用いる音声と手書き文字のずれのモデル学習に用いた. 学習セットを用いて求めた音声と手書き文字のずれ $\overline{T}$ および (2) 式の分布 $p(\delta)$ のパラメータの値を Table 1 に示す. これより, 音声よりも手書き文字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.

*A Study of Interface Having Simultaneous Inputs of Speech and Handwritten Characters. by NAKAGAWA Ryuta, KOBAYASHI Yui, KOBAYASHI Ryuji, SHINODA Koichi, and FURUI Sadaoki (Tokyo Institute of Technology)

Table 1 Time difference (δ) between speech and handwritten character inputs.

Subjective ID	0	1	2	3	4	5	6	7	8	9
# of characters	123	152	101	231	162	105	96	133	129	138
Average $\overline{T}$ (ms)	92.7	-77.8	44.5	-154.1	43.0	-17.2	51.4	-17.6	2.0	-14.7
Standard deviation (ms)	172.3	317.0	143.6	185.0	160.4	143.5	78.6	188.3	172.7	151.7

3.2 認識手法

手書き文字認識手法は，ストローク単位HMM[1],[2]を用いた．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．各HMMは，3 状態 1 混合とし，濁音，半濁音を含む平仮名 82 文字をオンライン手書き文字データベース [3] 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．

音響モデルは，2,000 状態 16 混合性別非依存 triphoneHMM 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ヵ月分の出現頻度上位 20,000 語をそれぞれ用いた [4]．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，単語間にショートポーズがある場合に対処するため，残りのすべての単語の先頭に，ショートポーズを許すスキップありの 1 状態無音 HMM を追加した．音響特徴量は 12 次元 MFCC とその差分，パワーの差分の計 25 次元を用いた．

3.3 実験結果

収録データのうち，学習セットに含まれない各被験者共通の 10 文 (形態素数 209)，計 100 文を評価セットとした．評価セットにおける入力手書き文字数は 412 文字で，手書き文字のみによる認識での平均認識率は 87.4%であった．ベースラインとなる音声のみによる認識 (SPCH) での単語正解精度は 60.8%であった．

音声認識と手書き文字認識の 2 パス処理による認識の結果を Fig. 2 に示す．ここで，手書き文字尤度重み α は被験者ごとに最適な値を採用している．手書き文字認識を加えたMTD1では，SPCHと比べ単語正解精度が 1.9 ポイント改善した．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を考慮したMTD2ではSPCHに比べ，2.2 ポイント改善した．これらより，提案するインタフェースが音声認識性能を改善し，音声と手書き文字の同期のずれを考慮することでさらに性能が改善することがことが確認された．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ALGNでは，MTD2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．

なお，手書き文字尤度の重み係数 α を全被験者で共通として実験を行ったところ，MTD1では，音声のみの結果を上回ることはなかった．またMTD2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った．

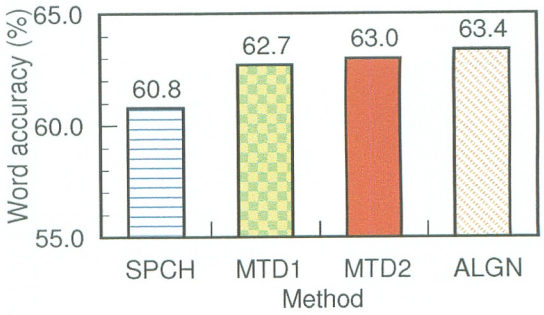

Fig. 2 Recognition accuracy of the proposed method.

4 おわりに

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．

また，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したい．

謝辞 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] 嵯峨山 他，信学会 PRMU-35, 1-8, 2000.
[2] 中井 他，信学会 PRMU-36, 9-16, 2000.
[3] 中川 他, 信学会 PRU-115, 43-48, 1995.
[4] IPA「日本語ディクテーション基本ソフトウェア 1999 年度版」
[5] 市屋 他，信学会総会 (春), D-14-007, 2004.
[6] 中川 他，情処 NL-167 SLP-56, 29-34, 2005.