論文 / 著書情報
Article / Book Information

論題(和文)	音声とペンの準同期入力に対するマルチモーダル認識
Title(English)	
著者(和文)	岩田 憲治, 渡邉 康司, 中川 竜太, 篠田浩一, 古井貞熙
Authors(English)	Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会 2006年秋季講演論文集, Vol. , No. , pp. 45-46
Citation(English)	, Vol. , No. , pp. 45-46
発行日 / Pub. date	2006, 9

音声とペンの準同期入力に対するマルチモーダル認識*

◎岩田憲治, 渡邉康司 (東工大), 中川竜太 (長崎大), 篠田浩一, 古井貞熙 (東工大)

1 はじめに

複数のモードを同時に認識することで，頑健な認識・理解を可能とするマルチモーダル認識の研究が，現在盛んに行われている．本研究では入力速度が異なり，入力が準同期的である連続音声とペン入力を複数モードとして用い，それらを同時認識するアルゴリズムを提案する．

2 同時入力インタフェース

データ収録では，タブレットパソコンを使用し，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，使用者が入力可能なタイミングでペン入力を行う．その際使用者は，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．このような条件を満たすペンの入力形態は様々なものが考えられる．本研究では [1] と同様，

1. 文節先頭文字の平仮名の入力
2. 文節先頭文字の平仮名の 1 画目のみの入力
3. 文節開始のタイミングの入力
4. 文字テーブルによる文節先頭文字の平仮名の「行」の選択

という操作を行う 4 種類の形態を用意し，それぞれについて評価を行う．それぞれの入力形態の特徴の比較を Table 1 に示す．

Table 1 各入力形態の比較

入力形態	先頭文字	1 画目	区切り	文字選択
1 回の入力時間	×	△	○	○
1 回の入力情報量	○	△	×	△
同期の取りやすさ	×	△	○	△

3 マルチモーダル認識

本研究では，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う．第 2 パスでは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．

3.1 第 1 パス

本研究では，音声の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するよう使用者に依頼しているが，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，ペン入力のずれ μ(ペン入力開始時刻 − 対応する単語の開始時刻) を設定し，μ だけペン入力開始時刻を前に補正する．ここで，μ は負の値も取り得る．

前述の補正の結果，ペン入力開始時刻は対応する文節発声中に存在するものとする．そして，ペン入力開始時刻に存在する音声を単語単位に分割したセグメントと，ペン入力のセグメントを対応づけ，ペン入力認識尤度を重み付けして音声認識の単語仮説に反映させる．この重み係数を α とする．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. 1 に示す．これにより，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．

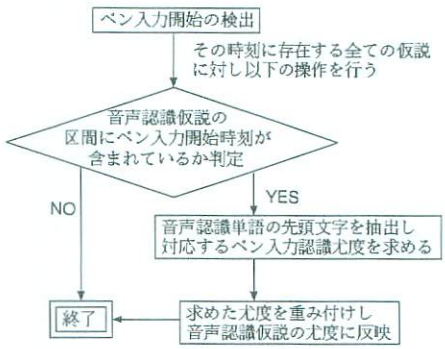

Fig. 1 ペン入力認識尤度を音声認識の単語仮説に反映させるアルゴリズム

3.2 第 2 パス

第 1 パスでは，ペン入力開始時刻に存在する全ての音声認識の単語仮説に対し，そのペン入力の尤度を反映させていた．しかし，開始時刻がペン入力開始時刻と近い単語仮説と離れている単語仮説があったとき，前者の仮説の方が正解である可能性が高いと考えられる．ここでは，音声とペン入力の同期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．

今，ペン入力 c_n の入力開始フレームを i_n，ある単語仮説 (アーク) の開始フレームを i としたとき，フレーム i_n と i との差を δ とする．δ の分布 $p(\delta)$ は正規分布に従うとする．このとき，このアークに対する重み係数を $\beta p(\delta)$ とする．ここで，β は重み係数の調整パラメータである．そしてペン入力の前後 I フレーム内に先頭フレームが存在するアークの尤度に対し，重み係数を用いた重み付けペン入力認識尤度を乗じ，新たな認識尤度とする．ここで I は定数である．単語グラフとペン入力の関係を Fig. 2 に示す．ペン入力認識尤度を加えるアークの先頭を黒丸で表した．

4 評価実験

前述のアルゴリズムを Julius[2] に実装し，認識実験を行った．

---

*Multimodal recognition for quasi-simultaneous speech and pen inputs
By Kenji Iwata, Yasushi Watanabe (Tokyo Institute of Technology), Ryuta Nakagawa (Nagasaki University), Koichi Shinoda and Sadaoki Furui (Tokyo Institute of Technology)

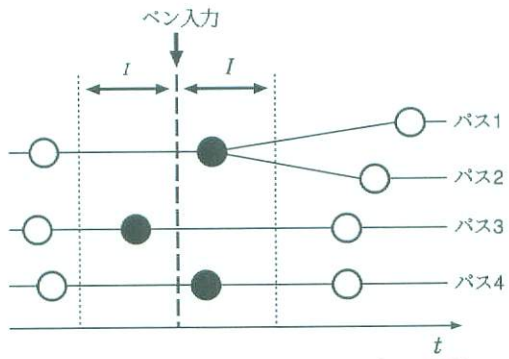

Fig. 2　単語グラフとペン入力の関係

4.1　評価データ

2 章で述べた 4 つの入力形態について，オフィス環境で日本人男性 10 名の収録を行った．入力形態ごとに自由発話 5 文と，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 15 文の計 20 文を入力した．この計 800 文に対して展示会場の雑音を SN 比 20dB，15dB，10dB で重畳させたもの，雑音を重畳してないものの 4 種類，計 3200 文を評価データとして用いた．また，被験者は収録前に各インタフェースをしばらく使用し，ある程度慣れた状態で収録を行った．

4.2　実験条件

音響モデルは情報処理振興事業協会 (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 triphone HMM を用いた [3]．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」[4] 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．また，評価データから被験者ごとに音声とペン入力のずれ (ペン入力開始時刻 − 対応する単語の開始時刻) の平均 μ を予め求めておき，認識を行った．

4.3　実験結果

各評価データで音声のみの認識をしたときの結果，および，音声とペン入力のマルチモーダル認識をしたときの結果を Table 2 に示す．第 1 パスの重み係数 α および第 2 パスの重み係数 β，ペン入力認識尤度を反映する範囲を定める I は，評価データに重畳させる雑音や入力形態において，それぞれ全被験者で共通とし，全体の認識性能が最も高くなるものを事後的に採用した．最適な α，β，I は雑音環境や入力形態によって異なったが，それぞれ α =0.001〜0.085，β =0.001〜0.08，I =10〜20 フレーム (1 フレーム=10msec) の範囲となった．全ての場合において音声とペン入力のマルチモーダル認識の単語正解精度が音声のみの認識の単語正解精度を上回るという結果を得た．

Table 2　認識結果 (単語正解精度 (%))

SN 比		先頭文字	1 画目	区切り	文字選択
雑音なし	音声のみ	83.6	84.0	84.5	82.6
	音声+ペン	84.6	84.4	84.9	84.5
	改善	1.0	0.4	0.4	1.9
20dB	音声のみ	76.8	76.7	79.2	76.5
	音声+ペン	78.5	77.1	80.0	79.0
	改善	1.7	0.4	0.8	2.5
15dB	音声のみ	67.1	63.4	66.8	66.8
	音声+ペン	68.3	64.0	67.7	69.7
	改善	1.2	0.6	0.9	2.9
10dB	音声のみ	41.6	41.0	43.0	42.0
	音声+ペン	44.4	41.8	43.8	46.0
	改善	2.8	0.8	0.8	4.0

入力形態で比較すると単語正解精度の改善の大きさは，ペンタップによる文字選択 > 文節先頭文字の入力 > 文節区切りの入力 ≥1 画目のみの入力，となった．これは 1 文あたりのペン入力の情報量の多さに加えて，音声との同期の取りやすさが認識性能に影響していると言える．また，雑音を重畳させた評価データでは，一部を除き，全ての入力形態において雑音の SN 比が大きくなるにつれ単語正解精度の改善が大きくなった．これは本研究のマルチモーダル手法が雑音下において特に有効であることを示している．

5　まとめ

本研究では，連続音声とペン入力を同時認識し，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ながら最適解を探索するアルゴリズムを提案した．ペン入力においては 4 種類の入力形態を用意したが，音声認識のみの結果と比較し，いずれの入力形態でも認識性能の改善を確認した．認識性能は 1 文あたりのペン入力の情報量と，音声との同期の取りやすさが影響していると言える．また，雑音下において特に有効な手法であることを示した．

本研究では，マルチモーダル認識での各パラメータは事後的に最適なものを採用したが，今後，それらの値を環境や話者の違いに対し自動的に適応させる手法の研究が必要である．また，今回は手書き文字の認識結果を予め求めていたが，認識デコーダ内で音声と手書き文字を同時に認識を行い，結果を統合するアルゴリズムを構築することが必要である．

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] 渡邉 他, 信学技報 SP-19, 49-54, 2006.
[2] 李, 情処研報 SLP-59, 127-132, 2005.
[3] 鹿野 他, “音声認識システム”, オーム社, 2001.
[4] 中川 他, 信学技報 PRU95-115, 43-48, 1995.