音声とオンライン手書き文字のマルチモーダル認識*

1N−05

白 銘† 長井 隆行‡ 榑松 明§
電気通信大学大学院 電気通信学研究科¶

1 はじめに

近年、コンピュータの普及により、新しいユーザ・インタフェースに注目が集まっている。音声認識装置や、ペン入力文字認識装置などが出回りつつある。実環境下での使用を考えた時、周囲環境に雑音が存在するために音声認識性能の低下が避けられないことがある。一方、手書き文字の認識も書き方が悪いと認識率の低下を生じる。本研究では、音声認識とオンライン手書き文字認識の情報を併用して、統合した認識結果を出す方法を示す。複数モードによる相補性により認識率により、入力作業の簡略化が図られる。

2 システムの構成

2.1 システムの構成

統合システムとしては、隠れマルコフモデル（HMM）を使用した音声認識とオンライン手書き文字認識の出力を統合し、統合した認識結果を出力するように構成する。

2.2 HMMによる音声認識

HMMは音声認識分野で最も有効な方法として用いられている。本研究はHMMによる、音素モデルの結合により音節モデル[1]を構築する。連結学習により、音節に適したパラメータ値に調整し、認識の際はViterbiアルゴリズムで尤度の高い順にN個の音節モデルを選択する。

2.3 HMMによる手書き文字認識

オンライン手書き文字認識の研究分野では、続け字、崩し字、自由書式の文字に関して、認識が難しいことが問題となっている。これは筆記者により時間軸方向の変動が大きいためである。HMMを用いることにより、変動に強くなることが考えられる[2]。

本研究では、オンラインで入力された手書き文字パターンの点をペンの変化方向で量子化する。16方向のベクトルに、裏ストロークベクトル(pen up時のベクトル）を加え、計32種類のベクトルによって入力パターンをベクトル列に置換する。モデルの作成の際には、方向のベクトル列と枠中心から各座標の距離の2種類の特徴ベクトルを用いて文字毎にHMMモデルを作成する。未知入力文字パターンは特徴ベクトル量子化を行い、各座標の中心からの距離を加えてモデルに入力する。この入力に対して、HMM辞書内の各モデル毎に尤度計算を行い、尤度の高い順にN個の文字モデルを選択する。

2.4 音声認識と文字認識の統合

音声と画像の統合では、特徴量のレベルで行う方法[3]と、音声のみによる認識、画像のみによる認識を別々に行い、両方の認識結果の情報を統合する方法がある[4]。音声のみによる認識と画像のみによる認識を分離した場合は、各モードの認識装置のパラメータが独立に調整可能なので、より高い認識率の達成が期待できる。本研究では、音声認識と手書き文字認識を別々に行い、両者認識結果である尤度を1次統合を用いた。まず、音声と文字と別々にすべての文字（音節）に対する尤度を計算しておく。尤度の高い順にN個候補の対数尤度を出す。次に、音声のN個候補の対数尤度と画像のN個候補の対数尤度を式(1)で統合する。音声認識と文字認識ではHMMを用いるが、ベクトルサイズの違いや尤度のばらつきの影響をなくすため、正規化を行う。最後に、同一の文字(音節）に対する高い順に新しいN個の候補対数尤度を得ることにより、認識率を評価する。

$$P_t = \alpha(P_s - \mu_s) + (1-\alpha)(P_w - \mu_w) \qquad 0 \le \alpha \le 1 \quad (1)$$

ここで、P_tは統合後の対数尤度、αは音声の重み、P_s、P_wはそれぞれ音声、文字の対数尤度、μ_s、μ_wはそれぞれ音声、文字の対数尤度の平均を表す。

3 実験

3.1 実験条件と方法

表1に実験条件を示す。文字認識実験では字体制限なし、ページ内に筆記枠があるという条件で50名筆記者により書かれた当研究室で収集した日本語片仮名データベースを用いた。40人のデータを用いて学習を行い、残り10人のデータを用いて認識実験を行った。

*Multi-modal Recognition of Speech And On-line Handwriting Character
†Ming Bai
‡Takayuki Nagai
§Akira Kurematsu
¶Graduate School of Electro-Communications, University of Electro-Communications

音声認識用のHMM音響モデルは研究室で作成した音節モデルを初期値として、音声情報処理研究用日本語音声データベースの男性3名音素バランス単語（216単語）データを用いて連結学習（8回）を行い、残り3人のデータを用いて音声認識実験を行った。雑音データとして、電子協雑音データベースに収集した人混みの雑音を用いた。

表 1: 実験条件

文字	特徴パラメータ:方向ラベル列(32ラベル) 枠中心からの距離 文字HMMモデル: 5状態3ループして徐々に増やしていき 1混合のleft-right型HMM 出力確率は連続確率分布でガウス分布
音声	音節HMMモデル: 5状態3ループ5混合のleft-right型HMM (1音素長の音節) 7状態5ループ5混合のleft-right型HMM (2音素長の音節) 出力確率は連続確率分布でガウス分布 特徴パラメータ:16次MFCC △16次MFCC 対数エネルギー △対数エネルギー 量子化ビット数：16bit 分析フレーム長：25.6ms 時間窓：ハミング窓 高域強調係数：0.95

文字認識のための前処理として、タブレットにペンを降ろすときに発生することが多い「連続する同一座標点」を入力パターンの座標点列からの除去および文字のサイズの正規化する。特徴ベクトルは方向ベクトル列と枠中心から各座標の距離の2次元で表わす。

音声認識の実験は、表1に示す条件によって、音節モデルを作成し、連結学習、認識を行う。音素バランス単語データベースは連続発声されたデータであるため、認識する時挿入誤り、脱落誤りになることがある。文字については、文字ごとに枠を設けるので、挿入誤り、脱落誤りになることがない。本研究では、文字側で得られる枠数により音節の数を制約することとする。式(1)の係数αを0.0から1.0まで変化させて最適な係数αを求める。

3.2 実験結果と考察

認識率は次の式を用いて求めた。

認識率＝（正解文字(音節)数／全文字(音節)数）＊100%

文字認識ではHMMの状態数を変えた場合第5候補までに正解がある認識率は最高で86.2%である。音声のみの第5候補までに正解がある認識率は89.8%である。雑音のない音声と人混み雑音を元の音声の平均パワー比で-10dbから5dbまで段階的に付加した音声を用いて、係数αを0.0から1.0まで変化させた場合の統合認識の結果を図1に示す。図1に示されるように、雑音のない音声と文字を統合した場合、最大で92.3%の認識率が得られた。雑音を加えて信号対雑音比-10dBの音声のみの認識率は69.5%であるが、統合した結果は88.5%の認識率となる。全ての信号対雑音比に渡って統合後の認識率が音声のみの認識率を上回っていることが図に見られる。特に雑音のパワーが大きくなるときに、認識率の低下する度合が統合後の方が小さいことが分かり、異なるモーダルの統合に効果的であると言える。

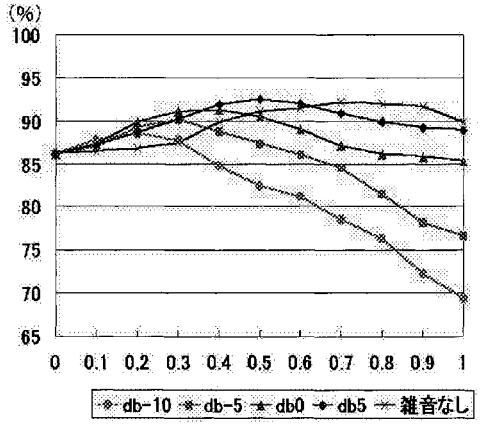

図 1: 異なる信号対雑音比における結合係数と認識率の関係

4 むすび

本稿では、音声とオンライン手書き文字を統合したマルチモーダル認識手法を示した。この方法により、雑音環境下でも認識性能が向上することを示した。今後の課題としては、ニューラルネットワークを用いるなど音声と文字の統合方法を更に検討することが挙げられる。

参考文献

[1] 川端崇裕,“音節言語モデルを用いた固有名詞の音声認識”電気通信大学2000年度修士論文,2000年.

[2] 伊藤等, 中川正樹,“Hidden Markov Modelに基づくオンライン手書き文字認識”信学技報IE97-54,PRMU97-85,MVE97-70(1997-07)

[3] 呉, 田村, 他,“音声、口形特徴量を併用するニューラルネットを用いた母音認識”信学論D-II,Vol.J73-D-II No.8,pp.1309-1314,1990

[4] 新谷, 萩原,“視聴覚融合によるHMM音声認識信学会春季大会”, A294,1994.