99-01069

バイモーダル情報による個人識別の研究

北　村　　正　　名古屋工業大学教授

１．研究の目的

近年，様々な社会システムにおいて，セキュリティの問題が重要な役割を占める場が多くなってきている。個人を機械により自動識別することもその一つである。現在，最も一般的な個人識別システムとして，磁気カードと暗証番号によるものがあるが，紛失や盗用などの問題がある。そのため，個人に固有な情報である指紋，顔，筆跡，音声などの情報を用いて個人識別を行うものが考えられている。このうち，音声や顔情報を利用する方法は，ユーザにとって自然かつ負担が少ない方法であるが，それぞれに背景雑音の問題，照明の変化などの問題があり，単独の手法によっては十分な結果を得ることは困難である。また，実環境下では，音声や顔などの必要な情報を正確に抽出するための技術が必要である。

一方，音声による個人識別には，個人の発声の時間変化情報が有効であることが知られているが，発声時報（音声，唇動画像）を組合せてバイモーダル情報とし，互いに欠点を補いあうことにより，より精密で実環境に強い個人識別システムを検討する。

２．音声情報を用いた個人モデル

2.1　データベースと個人モデル

本研究では音声と唇画像を取り扱うため，小規模のバイモーダルデータベースとして，研究用に利用可能なＭ２ＶＴＳ（話者37名，フランス語10単語，５時期）を用いた。

個人モデルは，唇画像の時間系列ベクトルの時間方向，話者間の変動を吸収できる隠れマルコフモデル（HMM）により作成する。音声，唇画像ともに５時期中の４時期分を学習データ，残り１時期をテストデータとし，時期の組み合わせを変えて実験を行う。テストデータ数は1850サンプルである。また，発声内容既知の個人識別を行う。

2.2　音声による個人識別実験

音声信号は，標本化周波数12kHz，フレーム周期10ms，フレーム長30ms ブラックマン窓でメルケプストラム分析を行った。特徴ベクトルとしては，１次から１２次までのメルケプストラムとフレーム間の１次差分ΔおよびΔの１次差分ΔΔの36次元のベクトルを用いた。上記の音声の特徴ベクトルを用いて，各個人の各単語ごとに，統計的手法に基づく隠れマルコフモデル（HMM）を作成する。HMMは連続確率分布の単混合のものであり，対角共分散を用いたleft-to-right モデルである。

以上の条件により，音声情報を用いた個人識別実験を行った。実験では，単語音声の始点と終点は既知としている。HMMの状態数を単語内の音素数により９～１５（音素数×３）とした場合の個人識別率を表2.1に示す。単語の個人モデルによる個人識別率は単語により異なるが，最低85.9%～最高98.4%であり，平均すると94.5%である。単語長が0.4秒前後と短いため，十分な識別率とはなっていないが，単語の組合せなどで単語長を長くすることにより識別率の向上が可能と考える。

表2.1　各単語の平均識別率（%）およびHMMの状態数

単　語	0	1	2	3	4	5	6	7	8	9	平均
識別率	97.8	95.1	93.0	98.4	85.9	97.3	95.7	95.7	88.1	98.4	94.5
状態数	12	6	9	12	15	9	9	9	9	9	

2.4 雑音による識別性能の劣化

音声のみで個人識別を行う場合，前述のように実環境では背景雑音の問題がある。ここでは，背景雑音の影響が識別率にどのような影響を与えるかを実験により検討した。実験では，ガウス性の白色雑音を単語音声に波形上で所定のＳＮ比となるように加算する。このときの単語音声の始点と終点は既知としている。実験結果を表2.2に示す。ただし，各単語のHMMの状態数はすべて８とした。

表2.2　SN比による個人識別率の影響

SN比	Clean	25dB	20dB	15dB	10dB
識別率	94.3 %	88.8%	75.8%	49.6%	26.0%

表に見られるように，背景雑音の影響はかなり大きく，37名の個人識別にもかかわらず，15dBで49.6%の識別率しか得ることができない。従来の単語音声認識実験によれば，単語音声認識でも背景雑音の影響をかなり受けるが，上記のように大きな認識率の劣化は見られない。従って，音声を個人識別に用いる場合は，音声認識以上に背景雑音の影響を緩和することが重要となる。

３．唇動画像情報を用いた個人認証モデルの検討

２章で述べたように，音声を用いた個人識別は背景雑音の影響を強く受ける。そこで，ここでは雑音の影響を受けない唇画像を用いた個人識別について述べる。

一般に物体を扱う画像処理法としては，モデルベース法とピクセルベース法（またはイメージベース法）がある。前者は，物体のモデルを仮定してそのモデルを推定し，そのモデルに基づいて物体の認識などを行う。しかし，適当なモデルが得られれば，照明条件の変化，物体の位置ずれなどに強いが，適当なモデルの仮説とその推定が困難なことが多い。後者は，画像の画素値をそのまま用いるものであり，処理が簡単という利点があるが，画像の照明条件の変化や物体の位置ずれなどの問題がある。本研究では，後者のピクセルベース法を用いるが，そのために照明条件の変化と位置ずれに対する対処法を検討した唇画像による個人モデルを提案している。

3.1 画像データのサブサンプリング

本研究では，歯，舌なども観測できるピクセルベース法を用いているが，唇画像の各フレームの画像サイズは80×40ピクセルと大きく，このままでは3200次元のベクトルを扱うこととなるため，唇画像を５×５のブロックサイズのサブサンプリングして16×8ピクセルとし128次元に情報圧縮したものを用いた。また，サブサンプリング後の唇画像データにより，各単語ごとのHMMを作成し個人識別実験を行った。画像のフレーム周期は40msであり，HMMのための特徴ベクトルとして，16×8 の輝度値データとそのフレーム間差分Δとその１次差分ΔΔの384 次元のベクトルを用いた。

結果を表3.1に示す。音声を用いた場合と比べると，個人識別率は平均65.6 %とかなり低いことが分かる。

表3.1　各単語とその唇動画像による個人識別率（%）

単　語	0	1	2	3	4	5	6	7	8	9	平均
識別率	67.0	68.1	54.6	80.0	62.2	61.6	56.8	67.0	71.4	67.0	65.6

3.2 輝度と位置の正規化

輝度の正規化

撮影条件の違いにより生じる輝度値の違いによる影響を除去するため，各単語を構成するフレーム全体の平均輝度を求め，その平均値を各フレームの各画素の輝度値から減算したものを用いる方法により，正規化を行うこととする。

正規化を施した場合の実験結果を表3.2に示す。正規化により，“１”以外のモデルでは誤りが減少しており，誤り平均が34.4%から26.0%に減少した。全体としては，24.4%の誤り改善率が得られ，輝度値の正規化の有効性が確認された。

表3.2　輝度値の正規化後の唇動画像による個人識別率（%）

単　語	0	1	2	3	4	5	6	7	8	9	平均
識別率	76.2	66.5	66.5	82.7	74.6	70.8	71.4	76.8	79.5	74.6	74.0

位置の正規化

唇は発声に伴い所定の画面内で移動し，極端な場合は画面から逸脱することがある。そのため，唇の動きに追随する必要がある。そのための位置の正規化の手順は以下の通りである。

まず最初のフレームの唇位置の位置を視察により検出し，唇位置がフレームの中央になるように切り出す。その後のフレーム関しては，画像をソーベールフィルタによりフィルタリング後，現在のフレームと次のフレームとの画像の正規化相互相関により，次フレームの唇位置を検出しトラッキングを行う。この処理により，発声などによる唇の移動にかかわらず，唇が画面のほぼ中心ある画像の時系列を得ることができる。

表3.3　位置正規化後の唇動画像による個人識別率（%）

単　語	0	1	2	3	4	5	6	7	8	9	平均
識別率	77.3	65.4	58.9	77.8	65.4	66.5	54.1	68.7	69.7	71.4	67.5

唇の位置の正規化を施したHMMを用いた場合の個人識別実験の結果を表3.3に示す。位置正規化により，平均の誤りが34.4%から32.5%に減少した。全体としては，6.1%の誤り改善率が得られ，位置の正規化の有効性が確認された。

輝度と位置の正規化の併用効果

更に両者の正規化を施した場合の実験結果を表3.4に示す。両者の正規化により，平均の誤りが34.4%から22.2%に減少した。全体としては，35.7%の誤り改善率が得られ，輝度値と位置の正規化の両者を用いることの有効性が確認された。また，同じデータベースを用いた他機関の報告よりもよい識別結果を得ている。

表3.4　輝度値と位置正規化後の唇動画像による個人識別率（%）

単　語	0	1	2	3	4	5	6	7	8	9	平均
識別率	83.8	69.7	74.6	82.2	76.2	76.2	75.1	82.7	79.5	77.8	77.8

4．バイモーダル情報の統合法の検討

２章と３章では，音声または唇動画像のみの個人モデルに基づく個人識別実験を行い，それぞれの特徴を検討した。ここでは，これらのモードの統合法について検討する。本研究では，２種類のモードによるHMMを用いていることから，統合法として，図4.1のような結果統合，初期統合，合成統合が考えられる。

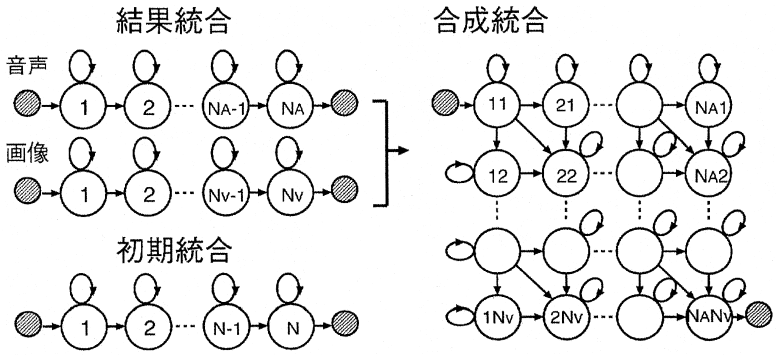

図4.1　バイモーダル情報の統合モデル

結果統合は，各単語に対する音声のHMMと唇動画像のHMMが個別に学習され，それらの出力結果である尤度を適当な結合係数により統合するものである。それぞれを独立に学習・認識するので，フレーム周期は異なっていてもよい。

初期統合は，モデル作成時において，HMMの各状態が適当な重み付けがされた音声と唇動画像の情報を持っており，一つのHMMが学習され，その出力である尤度を見るものである。この統合法では，音声と唇動画像のタイミングの同期がなされている必要がある。

合成統合は，音声HMMと唇動画像HMMの各状態の直積で表されるものである。この方法では，音声と唇動画像のタイミングの同期が必ずしもなされている必要はない。学習時の初期モデルとしては，それぞれを単独で学習した音声と唇動画像のHMMの各状態を適当な結合係数で重み付けをしたものを用いる。

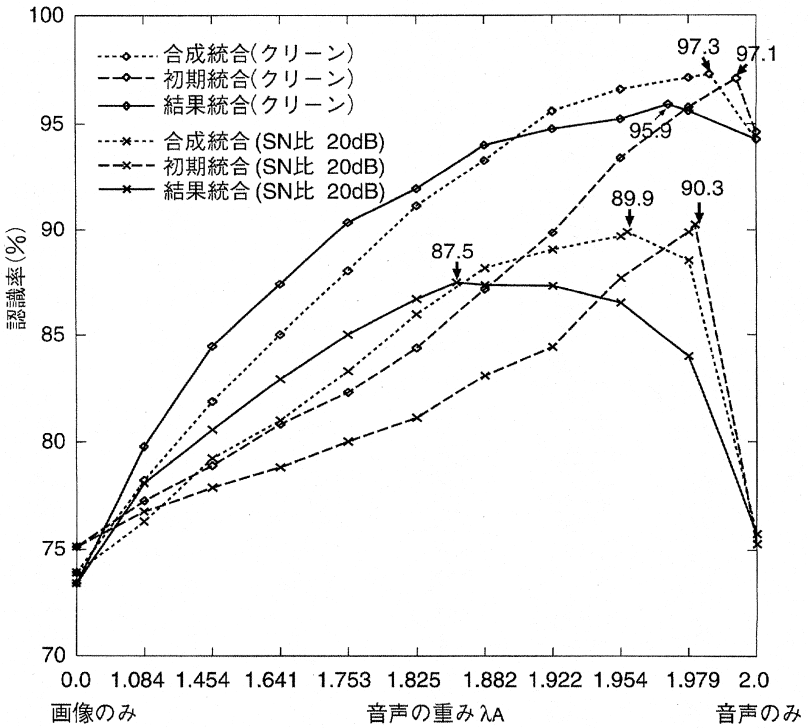

図4.2　雑音下での統合法と音声の重みによる識別率の関係

これら３つの統合法を用いた雑音下（クリーン，ＳＮ比15dB）での識別実験結果を図4.2に示す。図では，背景雑音の影響と統合の係数の重みを変化させた場合の識別率の変化を示している。HMMの状態数は，結果統合（音声８，唇動画像３），初期統合（８），合成統合（音声８×唇動画像３）とした。ただし，初期統合では，画像のフレーム周期が音声と同じとなるように，不足しているフレームは前後のフレームをコピーして用いた。

すべての結果において，画像のみ（重み0），または音声のみ（重み1）の識別よりも，音声と唇動画像のバイモーダル情報を用いることの有効性が確認できる。また，背景雑音のない場合（クリーン）では，初期統合（97.3%）と合成統合（97.1%）が結果統合（95.9%）よりもよい結果を示している。ノイズが混入した場合は同様な結果であるが，より画像よりの重みがよいことが分かる。ＳＮ比15dBの場合，音声のみの49.6%が81.4%と大幅に改善され，63.1%の誤り改善率となる。

音声と唇動画像の単語HMMの出力である尤度の値の統計によると，平均値では音声HMMが200 程度，唇動画像HMMでは−6000程度，標準偏差ではそれぞれ1300，16000程度と大きな差があることから，唇動画像HMMの出力確率が音声に比べてかなり小さく，音声情報に重きを置けばいいことが分かる。図4.2の実験結果は，これらの統計の結果と合致している。

むすび

本研究では，音声と唇動画像のバイモーダル情報を用いる個人識別について検討を行った。単語毎の個人の隠れマルコフモデル（HMM）により，個人モデルを作成した。Ｍ２ＶＴＳバイモーダルデータベース（10数字，37名，５回発声）による実験により，

（１）音声または唇動画像単独よりも，両者の組合せた方が識別率が向上する，

（２）背景雑音の影響に対して効果がある，

（３）合成統合，初期統合，結果統合の順に識別率がよい
など，提案法の有効性を確認した。

今回は比較的時間長の短い単語を用いたため，十分な識別率を得ることができなかったが，実用的な識別率を与える入力データ長の検討，さらにマルチモーダル情報を用いた個人認証について検討，大規模データベースＸＭ２ＶＴＳ（話者295名，英語文章＋10単語，４時期）を用いた個人識別実験などを行う予定である。

＜ 発　表　資　料 ＞

題　　名	掲載誌・学会名等	発表年月
音声と唇情報を用いたバイモーダル個人識別の統合法	電子情報通信学会総合大会 講演論文集　D-12-119, p.286.	2001年３月
音声と唇情報を用いたバイモーダル個人識別	電子情報通信学会ソサエティ大会 講演論文集　D-12-37, p.224.	2000年９月
視覚音声認識のためのコントラストの正規化	電子情報通信学会ソサエティ大会 講演論文集　D-12-31, p.218.	2000年９月