7N−5

セキュリティ管理の補助手段としてのニューラルネットを用いた話者認識

加藤 誠巳　　中條 有規
(上智大学理工学部)

1．まえがき

社会の情報化の進展に伴い、産業・経済等の情報に限らず、個人のプライバシーに関する情報も氾濫しつつある。そのため、個人のプライバシーを守り、犯罪防止にも役立てるため、「個人の同定」という課題について多くの研究がなされており、最近では、一般的な「暗証番号」による方法に代わる技術として、「指紋」・「網膜」・「顔」・「音声」を用いた手法が提案されている。なかでも、音声による話者同定の手法[1]-[3]は、その簡便性と、電話による通信手段の普及と相まって注目されている。今回、セキュリティールームの入室管理への応用を目標とした話者認識を、パーソナルコンピュータFM−R70HX2および、同社製ニューロボードNEUROSIM/Lを用いた階層型ニューラルネットワークにより実現したシステムについて検討を行ったのでご報告する。

2．音声データの仕様

今回の目的はセキュリティー管理であるため、微妙な個人差をも認識する能力が要求される。そこで、データを収集するにあたり個人差を小さくするため、以下のような通常よりも厳しい条件下で行った。

(1)発声者は、２０歳前後の男性のみとする。
(2)発音内容は、３秒間以内に発声された母音「あいうえお」とする。
(3)特異な発声法を避け、なるべく平静に発音する。

発音された音声を、サンプリング周波数12kHz、量子化ビット数16bitsのＰＣＭとし、分析に用いた。

今回の検討について、当研究室における入室チェックを想定し、男子学生12名のデータを識別すべきメンバーとして用い、他の男子学生15名を詐称者として選び、そのデータを利用して実験を行った。

3．入力音声の信号処理手法

収集した全ての音声データに対し、以下のような手順で信号処理を施した。

(1)無音部分の削除
経験的にしきい値を設定し、音声の有効部分の前後にある無音部分をデータ列から削除する。

(2)ケプストラム分析
データ列に窓長1024のハミング窓をかけ、窓を4分の1ずつずらしながらケプストラム分析を行う。
ケプストラム分析の具体的処理フローを図１に示す。

(3)ピッチ周波数抽出
男声のピッチ周波数の平均値mを125.0Hz、標準偏差σを20.5Hzとし、発声中の変動を考慮して、ケプストラム波形からm$\pm$6σの範囲内でのピークを検出し、その周波数をピッチ周波数とした。

(4)対数ケプストラムからのスペクトル包絡の抽出
ケプストラム波形の先頭から40点までにハミングリフタをかけ、ＦＦＴを施し、スペクトル包絡を得る。
包絡データは512点得られるが、データとして多すぎるので順に8個ずつ平均をとり、さらにその4個おきの、計16個の値を採用した。

(5)特徴点抽出
(3)・(4)ではそれぞれ分析窓数だけのデータが得られるので、発声時間によってデータ数が異なる。
時間軸上で、ピッチ周波数の一階差分、パワーの二階差分をそれぞれとり、それらの値の大きな点を発声内容の変化点（わたり）とみなして順次除去し、最終的に25ヶ所残し、データ数を均一にした。
従って、一回の発音に対し、（ピッチ周波数1個＋スペクトル包絡16個）×（時間軸上25点）＝425個のデータが得られる。

(6)データの正規化
ニューラルネットの入力には加工後のデータを0.0〜1.0の範囲に正規化して用いているが、単純な線形正規化の場合と、逆数をとって正規化した場合の二種類について検討を行った。

階層型ニューラルネットにおいて学習および認識実験に用いたデータは表１のとおりである。また、加工後の学習用スペクトル包絡データの例を図２に示す。ただし、詐称者のデータは学習には用いず、認識実験のみに利用している。

Speaker identification using neural nets for security management
Masami KATO, Yuki CHUJOH
Sophia University

4．階層型ニューラルネットワークによる学習と認識

入力データは、ピッチ周波数とスペクトル包絡であり、これらの異種ベクトルを一つのネットワークで学習するのは必ずしも適切とはいえない。そのため、入力層と中間層の間を全て結合したネットワーク（完全結合型）の他に、入力層と中間層の間を部分的に結合し、学習時における入力ベクトルの分離を施す手法（部分結合型）についても検討を行った。

今回の実験では、学習終了条件を「各カテゴリにおける教師データと発火値との差が全て0.05以下」とした。また、認識条件を「最大に発火したカテゴリの発火値が0.8以上であり、かつ、二番目に強く発火しているカテゴリの発火値との差が0.5以上」とし、認識条件に満たないものを詐称者とみなしている。

実験において、かなりよい認識率をあげたネットワークの例をそれぞれ図３に示す。また、このとき、各条件下における認識率は表２のとおりである。

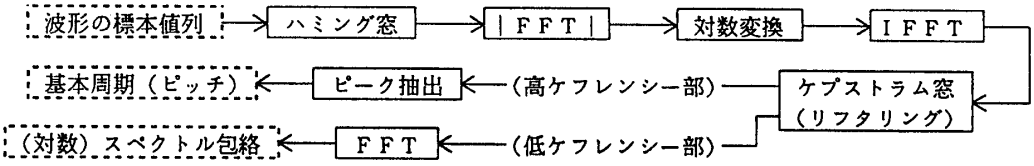
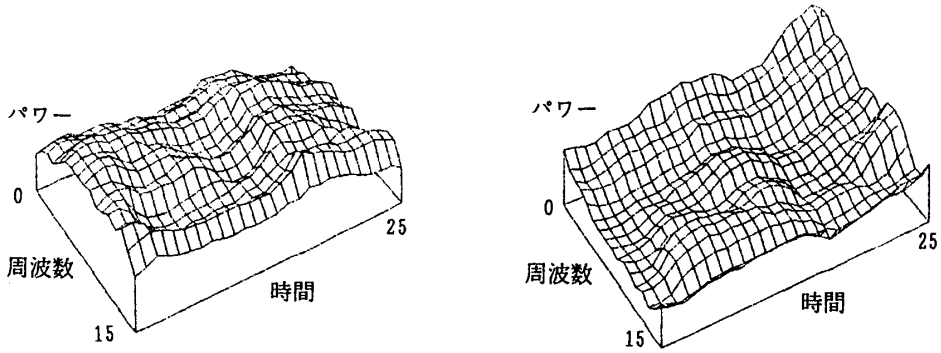
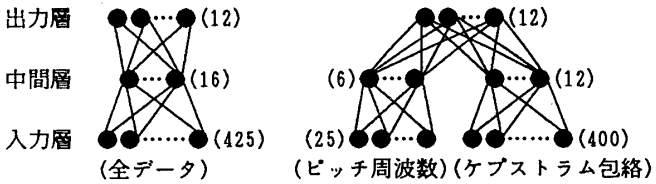


図１　ケプストラム分析と音声の特徴抽出の手順



(a)通常正規化　　(b)逆数正規化

図２　加工後のスペクトル包絡データの例



(a)完全結合型　　(b)部分結合型

図３　実験で用いたネットワーク

表１　実験に用いたデータ

	学習用	認識用	人数	合計
識別者	6個/人	4個/人	12人	120個
詐称者	なし	10個/人	15人	150個

（20歳前後の男性、「あいうえお」）

表２　ネットワークの構造と認識率

	ネットワーク構造(個)			認識率(%)			
				通常正規化		逆数正規化	
	入力層	中間層	出力層	識別者	詐称者	識別者	詐称者
完全結合型	425	16	12	87.96	88.24	87.96	79.55
部分結合型	25+400	6+12	12	85.88	94.36	91.13	84.16

5．むすび

現在、最も一般的なセキュリティーシステムは、銀行のＡＴＭ等で利用されている、ＩＤ番号（あるいはカード）と暗証番号（あるいはパスワード）を併用する方法であるが、ＩＤや暗証の貸与、盗用などの問題が起き易い。今後、音声などによる個人照合が容易に可能になれば、暗証番号に代わるチェック方法として広く利用されることが期待される。

最後に、有益な御討論を戴いた本学マルチメディア・ラボの諸氏に謝意を表する。

参考文献

[1] 古井：”ケプストラムの統計的特徴による話者認識”，信学論，J65-A，2，pp. 183-190（昭57）．
[2] 西村，海野，宮川，小池：”ニューラルネットワークによる話者認識(1)”，音響学会講演論文集，2-6-4，pp. 53-54（平1）．
[3] 野田，柳田：”ＨＭＭを用いた話者認識に関する検討”，音響学会講演論文集，2-3-5，pp. 59-60（平2）．