N５　正規型自然観測フィルタとニューラルネットワークによる聴覚機構のモデル化

高橋　秀岐†　堀　潤一††　齊藤　義明††　木竜　徹†

（新潟大学大学院自然科学研究科†　新潟大学工学部††）

１、はじめに

これまでに、音声認識を目的として人の聴覚機構のモデル化に関する研究が行われてきた。内耳の基底膜は、中心周波数が連続的に変化する帯域通過フィルタによって構成されている[1]。波形の再構成法として提案された自然観測システム[2]の構成が基底膜に似ていることから、自然観測システムを用いれば音声認識の改善が期待される。本研究では、正規型自然観測システム[3]によって基底膜のモデル化を行い、母音の認識率によって評価を行う。そして、信号の行列化が蝸牛殻神経での音波に対する反応と、ニューラルネットワークが聴覚神経路及び脳と、それぞれ対応する母音の音声認識システムを構築することを目的とする。

２、正規型自然観測フィルタ

正規型自然観測フィルタの出力波形は次式によって表わされる。

$$N_m^{(M)}(t) = {}_M C_m \Gamma^{M-m} \Lambda^m f(t) \qquad (m = 1,2,\cdots,M) \tag{1}$$

ここで、$f(t)$は入力波形、Mは項数、${}_M C_m$はM個からm個取る組み合わせ、Γ、Λはそれぞれ１次のLPF、HPFである。このとき、原信号$f(t)$は

$$f(t) = \sum_{m=0}^{M} N_m^{(M)}(t) \tag{2}$$

によって再構成される。図１に$N_m^{(M)}(t)$の周波数特性を示す。

３、信号の行列化

正規型自然観測フィルタの出力波形$N_m^{(M)}(t)$を、ニューラルネットワークに入力するために、一つのまとまったマトリクスデータに変換する。その手順を以下に示す。

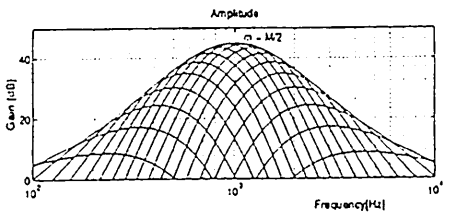

図１　周波数特性

1)　$N_m^{(M)}(t)$の絶対値をとる。
2)　絶対値化された波形をある一定の時間間隔で分割する。
3)　その区間毎の平均値を求める。
4)　平均値の最大値が１となるように規格化する。

1)～4)の過程より作り出されたマトリクスデータが、ニューラルネットワークの入力層になる。

４、ニューラルネットワークについて

聴覚神経路及び脳をモデル化するため、４層の階層型ニューラルネットワークを用い、教師付き学習を行った。それぞれの層をつなぐユニット間結合は、周波数領域では全結合型神経回路（ACNN: All Connect Neural Network）[4]を、時間領域では時間遅れ神経回路（TDNN: Time Delay Neural Network）を使用した。ACNNは、あるユニットに対して、前の層の全てのユニットが結合するニューラルネットワークである。TDNNは、時間軸上のあるユニットに対して前の層のある時間間隔で区切られたユニットだけが結合するニューラルネットワークである。TDNNは時間的なずれに強いという特性を持つ。今回使用したニューラルネットワークの概略図を図２に示す。

５、実験条件

音声データ及び認識システムの仕様を示す。

1)　音声データ

ATRにおいてサンプリング周波数f_s=20kHzで採取された男性話者１名の音声データを使用する。母音80個及び子音の後続母音40個から切り出し

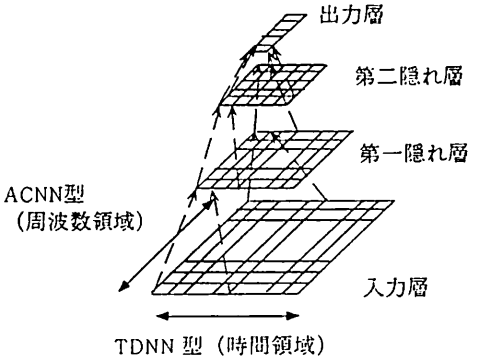

図２　ニューラルネットワークの概略図

た１ビッチ10ms、200ポイントのデータ120個を学習と認識に半分ずつ使用する。学習に際して、母音5種類を1セットとして12セット60個を繰り返し使用する。

2) 正規型自然観測フィルタ

カットオフ周波数f_c=1kHz、次数M=20とする。

3) 信号の行列化

周波数領域は21区間に、時間領域は1区間0.5msとし20区間に分割する。その結果、21×20のマトリクスデータが得られる。

4) ニューラルネットワーク

ユニット数は、入力層21×20個、第一隠れ層11×11個、第二隠れ層5×5個、出力層5個とする。教師信号は、5個のユニットがそれぞれの母音に対応した0、1の2値データとする。出力層のユニットには0～1の値が出力され、入力信号に対応する母音のユニットの値が0.95以上、他のユニットの値が0.01以下の場合、可認識とする。

6、結果

実験の結果、約7000回の学習で収束した。出力ユニットの例を表１に示す。また、同様な実験をバンドパスフィルタでも行った。その結果、正規型自然観測システムによる認識も、BPFによる認識も60種類の信号を用いた結果100%認識できた。

7、考察

従来から行われてきた音声認識では、音声データにフーリエ変換を施したマトリクスデータに対して教師付き階層型ニューラルネットワークが用いられてきた。本研究ではフーリエ変換を用いずに、フィルタリングによって前処理を行っているため、より人の聴覚機構に近似したシステムであると考えられる。聴覚フィルタ[1]は中心周波数に対する帯域幅が図３のように、高い周波数領域では帯域幅が広く、低くなるにしたがって帯域幅が狭くなるという特性を持つ。正規型自然観測フィルタはこの特性を維持しているが、聴覚フィルタに比べて全体的に帯域幅が広くなっている。式

表１ "あ"の信号を出力ユニットの例

あ	0.956834
い	0.001121
う	0.007643
え	0.002311
お	0.001232

(2)のMを大きくとればこの問題は解決できるが、M=30～40が数値計算上の限度であった。また、聴覚フィルタのエンベロープは低域ではなだらかで、高域では急峻な特性である。しかし、図１の正規型自然観測フィルタでは高い周波数領域で逆のエンベロープ特性を持つ。今回は特定話者の母音のみの認識対象としたが、不特定話者の子音を対象とする場合、より聴覚機構に近いモデルを考える必要がある。

8、まとめ

正規型自然観測フィルタとニューラルネットワークを用いて聴覚機構のモデル化を行い、母音と子音の後続母音を対象として音声認識の実験を行った結果、100%の認識率が得られた。今後は、子音及び不特定話者の認識を行う予定である。

参考文献

[1]赤木正人:信学論J77,9,pp.948-956(1994-9)

[2]飯島泰蔵:信学論(A)J76-A,11,pp.1620-1626(1993-11)

[3]滝田順子,飯島泰蔵,赤木正人:,信学技報,sp93-150,pp.46-54(1994-4)

[4]中野晏,飯沼一元,桐谷滋,他:" 入門と演習ニューロコンピュータ",技術評論社(1991)

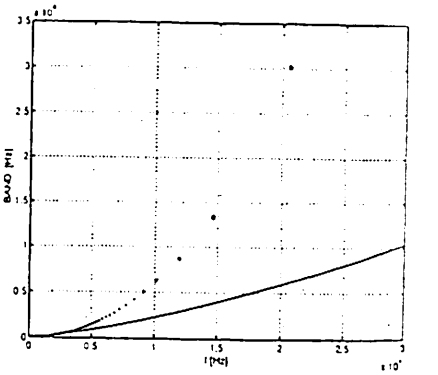

図３ 中心周波数－帯域幅特性