7F-3　時間遅れ要素を有するニューラルネットによる音声信号の時間構造の抽出

加藤誠巳　高木啓三郎　鵜飼敏之
(上智大学理工学部)

1. まえがき

近年、様々な分野において、ニューラルネットが注目され、成果を挙げつつある。特に音声の各分野でもその応用が試みられている。準定常的な母音部の認識においては、従来の単純なネットワークで十分であると考えられるが、実際の連続音声認識等への拡張を考えた場合、子音から母音への時間的変化や、母音間の調音結合等、時間的に徐々に特徴パターンが変化するため、それらの時間構造を考慮することが必要と考えられる。従来の単純なネットワークではこれら時間構造の抽出が困難であるため、種々の方法が提案されている。ここでは中間層に時間遅れ要素を追加することで、時間構造の抽出を行った結果について御報告する。

2. ネットワークの構造

従来の単純なネットワーク(図1)では、音声の時間変化を捉えることは極めて困難である。また時間構造を捉えるために入力層と中間層のノード数を増やすと(図2)、学習に時間がかかるうえ、局部極小点に落ち込む可能性も出てくる。

この様なことから、簡易でかつ拡張性のあるネットワーク構造を実現する必要がある。ここでは、時間構造を抽出するために、図3に示すような三層構造のネットワークを用いる。中間層の数は一層で、ニューロンの入出力関数としてはsigmoid関数を用いる。このネットワークは、図4に示すフレームkにおいて入力が与えられると、まず入力層から中間層へは信号が加えられるが、中間層の出力は最初は出力層へは伝えられず、時間遅れ要素(図3の□)に記憶される。次に入力層にフレームk+1の入力が与えられ、中間層に伝えられると、その中間層からの出力と、この直前に時間遅れ要素に記憶され

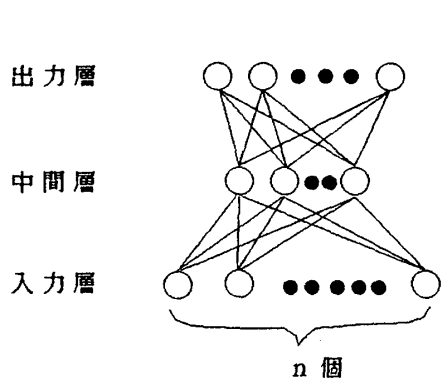

図1　従来のネットワーク

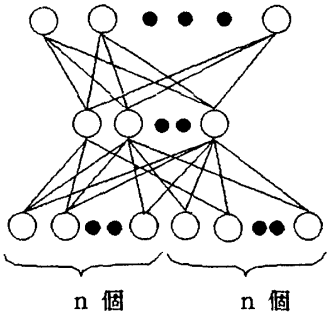

図2　従来の時間構造抽出ネットワーク

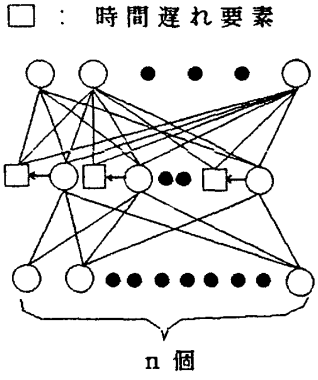

図3　時間遅れ要素を有するネットワーク

A Time Structure Extraction Method for Speech Signals by means of Neural Nets with Time Delay Elements
Masami KATO, Keizaburo TAKAGI, Toshiyuki UKAI
Sophia University

ていた出力が同時に出力層に伝えられ、出力層から結果が出力されるような構造を持つ。

この様に、ある時間に対してその直前の時間の中間層の出力を記憶させ、そのフレームの中間層の出力と共に出力層に伝えることにより、音声の時間変化を考慮にいれることの出来る構造になっている。

3. 時間構造抽出例

ここで提案したニューラルネットに0次を除く、20次までの最大値で正規化された低次のケプストラムデータを入力として加え、時間構造の抽出・認識を行った。ケプストラムデータの構造は図4に示すように10kHzサンプルされた音声信号の12.8ms窓長のデータを1フレームとし、相続くフレーム対のケプストラムデータを作り、学習に使用する。

学習の際には前半のケプストラムデータによる中間層の出力を時間遅れ要素に記憶したものと、後半のケプストラムデータによる中間層の出力によって、出力層の出力値が決まる。その出力値に応じて通常のバックプロパゲーションを行い、各リンクの重み係数、並びに閾値の更新を行う。このようにして学習させたネットワークを用いて、特定話者の5母音と母音間のわたりの6つの識別実験を行った。この場合母音間のわたりとしては「あ」から「い」、「い」から「う」、「う」から「え」へのわたりのみを対象とし、これらを「母音間のわたり」という1つのカテゴリーとした。

n=20として、従来型ネットワーク(図1)に対し後半のケプストラムデータのみを入力として加えたものと、n=20としたここで提案した時間遅れ要素を有するネットワーク(図3)によって行った認識実験の結果を以下に述べる。まず基礎実験として学習データに5母音の準定常部のみを用い、両ネットワークでclosed実験での認識率を100.0%にしたものでopen実験を行った結果、正解率は各々、99.4%、99.7%となった。

次に学習データとして、5母音の他、母音間のわたりのデータを学習に使い、各々のネットワークに対して認識率が99.0%を越えるまで学習させたネットワークを用いてopen実験を行った。その結果、従来型ネットワークで準定常部分、及びわたりの部分の正解率が各々92.3%、81.3%であったのに対し、時間遅れ要素を有するネットワークでは各々94.0%、90.1%となった。

この結果より、準定常部分のみの認識に関して両ネットワークは、ほぼ同等の性能を示すが、非定常なわたりの部分の認識に関しては提案したネットワーク構造を用いることにより、かなりの改善がみられた。

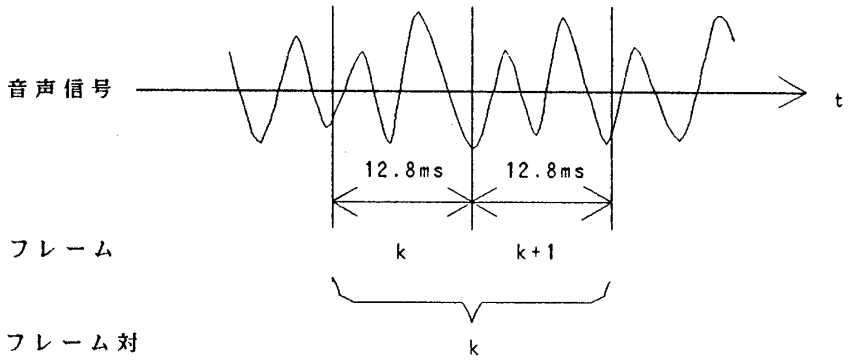

図4　音声データの切り出し

4. むすび

音声認識において要求される時間構造の抽出が容易に行えると考えられるネットワーク構造を提案し、母音間のわたりの認識に適用し、その有効性を確認した。本手法は時間遅れ要素の段数を増やすことによって容易に時間変化に対して拡張できるので子音の認識にも効果を発揮することが期待できる。