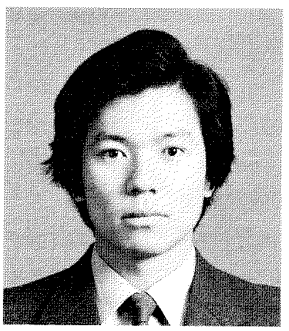
—平成元年度助成研究報告—

ニューラルネットに基づく音声情報圧縮



成蹊大学工学部
助教授 森島 繁生

1．はじめに

バックプロパゲーションを学習規則として用いた多層パーセプトロンは、パターン認識や信号処理など多くの分野で注目され、活発に研究が行われている。特に、非線形関数としてシグモイドが一般的に用いられており、この非線形性を有効に利用できるような応用分野で高い成果が報告されている。

ところで同一層内にリンクが存在しない多層パーセプトロンにおいて、中間層のユニット数を入出力層のユニット数よりも減少させることにより、中間層において入力信号の情報圧縮が可能となる。我々は特に音声信号を対象として3層型のニューラルネットの情報圧縮性能について検討を進めてきた。

2．3層型のシステム構成

図1に本稿で検討対象とするニューラル情報圧縮システムを示す。入力層と出力層のユニット数が分析フレーム長に相当する。ニューラルネットが最適な圧縮アルゴリズムを自動的に習得することを期待し、入力データおよび学習教師データとして、8 kHz、12ビットでAD変換した音声信号サンプルをそのまま用いることとした。

入出力層のユニット数と中間ユニット数を任意に設定して、フレーム周期と圧縮率をコントロールする。また、中間層ユニットの出力値を量子化することによって音声符号化が可能となる。学習方式はバックプロパゲーションを用いており、非線形関数として定義域が（－1、1）のシグモイドを用いている。学習データはある話者1名の発声した音声系列から切り出したものである。全パターン数は5440個で、この全パターンを順に入力した場合の学習回数を1回とカウントして、500回の学習を行った。収束は約250回でほぼ完了する。

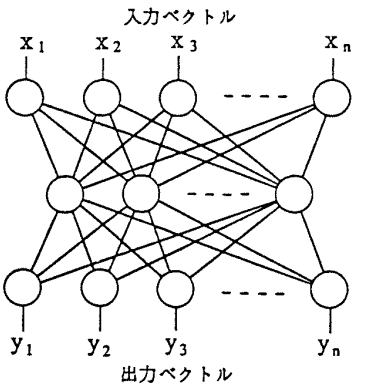


図1　3層ニューラルネット

3．圧縮率と SN 比の関係

学習を完了した3層ネットワークの性能評価を試みた。学習の際の重み係数の初期値は乱数によって与えた。入出力層のユニット数を一定とし、中間層ユニット数を変化させて学習し、学習話者と学習外話者2名の音声を入力して SN 比を測定した。ここで中間層の量子化は行っていない。図2は入出力ユニット数が8の場合であり、横軸に中間層ユニット数、縦軸に SN 比を示している。ここで SN 比は中間層ユニット数すなわち圧縮率と線形な関係にあることがわかる。また、学習外話者に関しては SN 比の上では、僅かに性能が劣化しているが、聞き比べによる主観評価ではその差は殆ど区別できない程度であった。この結果により恒等写像を実現する3層ネットワークは、学習データに強く依存しない音声に対して一般的な内部構造を学習したことになる。また、3名の話者で実験を行なった際の中間層各ユニットの出力分布を調べたところ、殆どゼロ付近に分布が集中しており、恒等写像を実現しようとする3層ニューラルネットは大部分シグモイドの線形部分を使用していることが分かった。

4．音声符号化の実現

入力層から中間層までを符号器、中間層から出力層までを復号器として、中間層出力を量子化することによって音声波形符号化が実現できる。あるビットレートを実現するには中間層ユニット数とビット配分のバランスが重要である。図3は入出力ユニット数が8の場合のビットレートと SN 比の関係を表わしている。横軸は圧縮の比率であり、プロットした点は各中間層出力に割り当てられるビット数を示している。さらに破線で便宜上の同一ビットレートの位置を示す。実線は各圧縮率で実現しうる SN 比の最大値（量子化なし）を示す。ここで各ユニットの出力は均等にビット配分されるものとする。これより、各ビットレートを実現するのに最適な圧縮比とビット数が存在することが分かる。例えば分析長が8サンプルで16 kbps を実現するには、中間層ユニット数を4とし、4ビットで量子化すればよく、SN 比は約15 dB と読み取れる。なお、ここでは各中間層出力値を別々に非線形量子化しており、学習時には量子化は行なっていない。

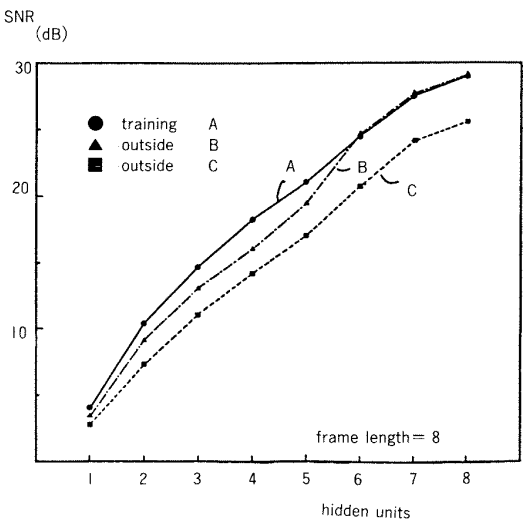


図2　圧縮率と SN 比の関係

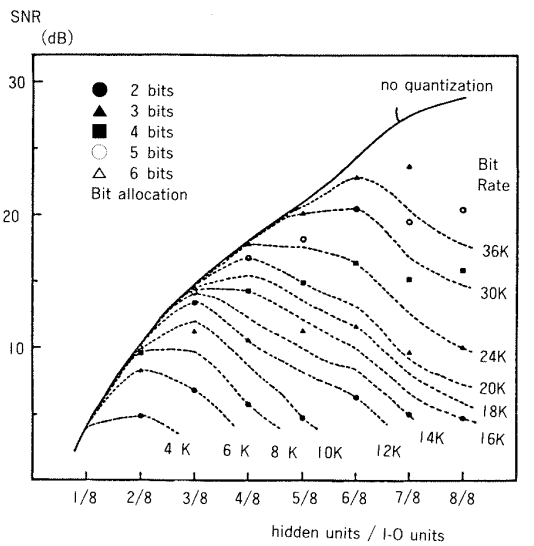


図3　ビットレートと SN 比

5．3層ネットの特性評価

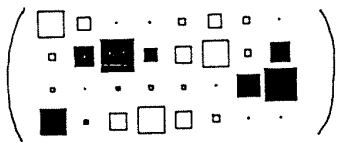
学習を完了した各層間の結合係数の分析を行なった。図4は入出力ユニット数8、中間層ユニット数4の場合にバックプロパゲーションによって形成された結合重み係数であり、入力ベクトルから中間層、中間層から出力ベクトルへの変換行列の形式で表示している。中間層出力はほぼゼロ付近に分布が集中していることから、入力から中間層、中間層から出力への写像は線形変換を仮定して問題ない。変換Aは入力層から中間層への結合重み、変換Bは中間層から出力層への結合重みであり、黒が正数、白が負数を表わし、正方形の大きさが係数の絶対値に対応している。また、BAが入力から出力への恒等写像を表わす8行8列の正方行列である。

Aと'Bとは、ほぼ類似した行列となっており、ABは対角行列となっている。したがって、砂時計形の3層ニューラルネットは恒等写像を実現するような学習法によって直交変換的な内部構造を学習したことになる。しかし、ここでの結果は線形空間内で説明のつくものでありシグモイドの非線形性を有効に活用しているとは言えない。すなわちこれは主成分分析によって入力ベクトル次元数よりも少ない次元の線形空間上に入力をマッピングして、再び元の次元に復元するやり方と大きな差はなく、非線形関数を用いている分、性能がこの主成分分析法よりも劣ることが予想される。

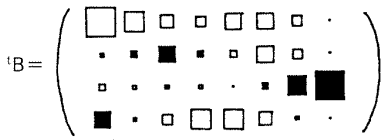
6．中間層 VQ の意義

入出力層8ユニット、中間層4ユニットとして学習し、同じ学習データを用いてその中間層出力を求め、これをトレーニング系列として4次元のベクトル量子化を行なった。入力信号のベクトル次元数は8であり、コードブックサイズの千倍がビットレートに対応している。この様子を表1に示した。

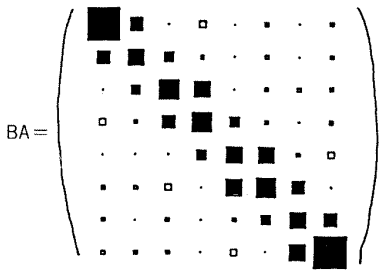
また、比較のため原信号サンプルを直接ベクトル量子化した結果も求めた。学習話者とはバックプロパゲーションおよびクラスタリングに用いた学習用音声そのもので検証した場合、学習外話者Aは男性、学習外話者Bは女性である。



a) Weight from Input to Hidden Layer



b) Weight from Hidden to Output Layer



C) Transformation from Input to Output

図4　8－4－8の場合の結合重み係数

(単位 dB)

コードブックサイズ	学習話者		学習外A		学習外B	
	直接	中間	直接	中間	直接	中間
5bit	9.28	9.32	7.44	6.88	6.32	6.00
6bit	11.08	11.15	8.57	8.22	7.36	7.17
7bit	12.67	12.67	9.71	9.31	8.45	7.97

表1　直接 VQ と中間層 VQ の性能比較

直接 VQ と中間層 VQ との性能は SN 比でほぼ同様の性能を示しておりベクトル次元数の削減に役立つことが分かる。

7．5層ネットの可能性

中間層のユニット数が十分大きな3層パーセプトロンを用いれば任意の連続なマッピングを実現可能なので、圧縮と伸長にそれぞれ3層を利用して、第3層のユニット数を入出力ユニット数よりも少なくした5層ネットワークを用いれば非線形性を有効に利用した情報圧縮の可能性がある。ここで圧縮空間上での線形性を保存するため第3層は線形ユニットとし、非線形の効果が期待できる第2層と第4層のみ非線形ユニットとした。

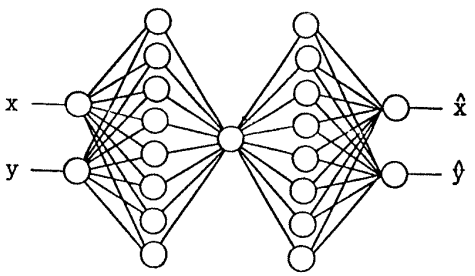


図5　5層ニューラルネットの構造

ここでは、とりあえず5層構造の性能評価を行なうため、図5に示すような簡単なネットワークを仮定して実験を行なった。2次元空間上の8つの点をランダムに選びだし、バックプロパゲーションによって学習を行なった。入出力は2次元、第3層は1次元とし、第2層と第4層は学習サンプル数と同じ8次元とした。学習回数は1万回である。図6は2次元空間上の学習に用いた8つの点とその再現点、さらに中間層出力として－1から1までの値を与えた場合の2次元の出力の再現点の軌跡を描いたものである。図7では今度は2次元空間上の全ての点を入力として中間層出力を調べ、同一の出力値を得る領域を等高線として示したものである。等高線の間隔は中間層換算で0.1であり、黒線の太さが0.02となっている。この5層ネットワークの変換により2次元空間上の全ての点は再現曲線上にマッピングされる。比較の

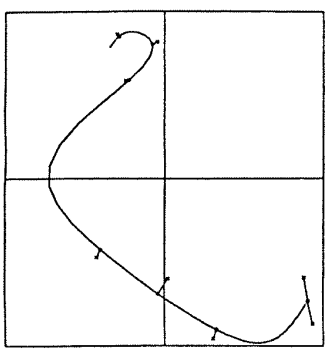


図6　5層ネットによる再現点とその軌跡

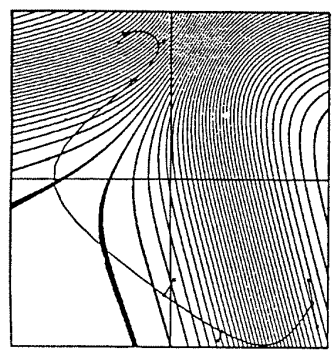


図7　5層ネットによる等高線

ため3層ネットワークの入出力層ユニット数2、中間層ユニット数1で線形関数を用いて学習を行なった。図8はこれにより実現される再現線を示しており、図9は同様にして求めた等高線である。すでに述べたように3層の場合には1次元の線形空間上へのマッピングに過ぎないことが見て取れる。これに対して5層の場合には、再現曲線が全ての学習点を結ぶように引かれており、3層の場合に比べて予測誤差が極めて小さいことが分かる。したがって、第2層と第4層の非線形ユニット数を適当に選び、第5層の線形ユニットの次元数も経験的に定めることにより、高性能な情報圧縮手法を見いだせる可能性がある。

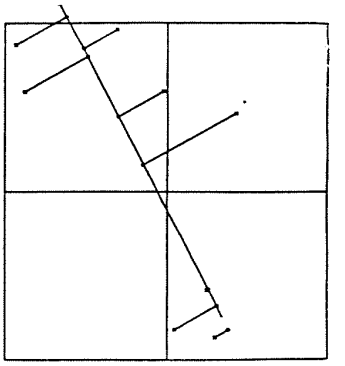


図8　3層ネットによる再現点とその軌跡

8．まとめ

3層ニューラルネットによって音声信号を教師データとして与え、恒等写像を実現するようにバックプロパゲーションによって学習させたところ、非線形性は生かされておらず、中間層ユニット数で決まる次元数の直交空間への線形変換的性質を学習していることが分かった。5層の場合には非線形性を生かした有効な情報圧縮の可能性があることを確認した。現在、実際の音声データを使って学習し、音声符号化器の性能評価を進めている。

最後にこの研究テーマに対して研究助成を頂き、研究の機会を与えて下さったサウンド技術振興財団に対し心より感謝の意を表します。

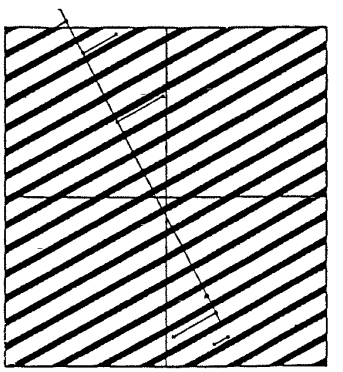


図9　3層ネットによる等高線