グラフィクスとCAD 48－9
（1990. 12. 14）

ニューラルネットによる母音認識における教師信号の検討

好田 正紀

山形大学 工学部

あらまし 多層パーセプトロン型のニューラルネットによる母音認識の学習において、入力音声の性質を考慮した教師信号設定法を検討した．その結果、ホルマントの母音5角形に基づいて教師信号を設定し、母音スペクトル間の距離の相対的な関係をできるだけ反映したマッピングは、より少ない出力ユニット数で、他の教師信号を用いる場合と同等の認識率が得られることを示した．このことから、一般に、入力層のユニットがはる空間での入力パターンの特徴ベクトルの分布と、出力層のユニットがはる空間での教師信号ベクトルの間のマッピングがより自然となるように、言い換えれば、入力層と出力層の間のマッピングの規則性が抽出しやすいように、カテゴリ間の距離の相対的な関係に関する事前の知識を考慮して教師信号を設定することが、ニューラルネットの学習を能率良く行う上で重要であるといえよう．

A study on supervisory signals in vowel recognition using artificial neural networks

Masaki Kohda

Faculty of Engineering, Yamagata University

4-3-16, Zyounan, Yonezawa-shi, 992 Japan

Abstract We investigate a vowel recognition based on artificial neural networks. In the training algorithm, called backpropagation, a supervisory signal should be carefully set to an appropriate value because it is one of key factors for carrying out the training algorithm efficiently and obtaining a high recognition performance. It was shown, in this paper, that the supervisory signal set to a typical value of the first and second formants for each vowel is superior to the conventional one because of reflecting spectral distances among vowels fairly well. This suggests that, in the setting of supervisory signals, it should be considered that the extraction of rules for nonlinearly mapping of the input pattern vectors scattered in a space spaned by units of input layer to the supervisory signals set in a space spaned by units of output layer becomes as easy as possible.

1. まえがき

近年、多層パーセプトロン型のニューラルネットにおける重み係数の学習に誤差逆伝搬法の有用性が示され、それを契機としてニューラルネットが種々の分野で利用されている[1]. 音声認識の分野でも、従来のＤＰ（dynamic programming）マッチングやＨＭＭ（hidden Markov model）による認識法のオルターナティブとして、ニューラルネットによる認識法が数年前から盛んに研究されるようになった[2].

ニューラルネットによる音声認識では、ニューラルネット自体の各種手法を単に利用するにとどまるのではなくて、ニューラルネットの構造・学習の中に音声特有の性質・知識を積極的に取り入れることによってはじめて、従来の認識法を超える可能性がでてくる.

ニューラルネットによる音声認識において音声特有の性質・知識を考慮したものをいくつか上げると、(a) 音声がもともと時系列信号であることを反映させるために、ニューラルネットを時系列構造化したWaibel等のＴＤＮＮ（time delay neural network）[3]、(b) 音声のスペクトル自体の変動と時間構造の変動の両方を考慮することができるように、ニューラルネットとＤＰマッチングを融合したSakoe等のＤＮＮ（dynamic programming neural network）[4]、(c) 音声のスペクトルの変動にその直前数フレームの影響を考慮することができるように、ニューラルネットを識別器としてではなくて予測器として用いた磯等のＮＰＭ（neural prediction model）[5]や、Tebelskis 等のＬＰＮＮ（linked predictive neural network）[6]、(d) 音声の調音結合を考慮することができるように、ニューラルネットの入力層のユニットに前後の音素の情報も入力するLeung等の研究[7]、(e) 音声のスペクトルの特徴を反映させたニューラルネットの重み係数の初期値設定法に関する入野等のＭＬＭ（multiple logistic model）[8]、等がある.

本稿では、多層パーセプトロン型のニューラルネットによる母音認識の学習において、入力音声の性質を考慮した教師信号設定法を検討する.

これまでは、入力のカテゴリ数と同数のユニットを出力層に並べて、正解ユニットの教師信号を１、他のユニットの教師信号を０に設定してニューラルネットの重み係数の学習を行うことが多い. この教師信号設定法は一見自然なようにみえるが、ニューラルネットによる処理はもともと、入力層のユニットがはる空間での入力パターンの特徴ベクトルの分布と、出力層のユニットがはる空間での教師信号ベクトルの間の非線形なマッピングを実現することであり、それを考えると１、０の教師信号は安易な設定法である.

ここでは、ニューラルネットによる母音認識における教師信号として、上記のマッピングがより自然になるように、言い換えれば、入力層と出力層の間のマッピングの規則性が抽出しやすいように教師信号を設定することが、ニューラルネットの学習を能率良く行う上で重要であることを示す.

2. 音声資料、分析方法

2.1 音声資料

男女各１名（ito、kat）が発声した 216語中の母音を音声資料とする. 母音の個数は、男声では 569個（/a/=161, /i/=105, /u/=116, /e/=79, /o/=108）、女声では 568個（/a/=161, /i/=105, /u/=115, /e/=79, /o/=108）である. 学習サンプルとして男女声とも各母音50個の計250個を用い、 残り（319個、または、318個）をテストサンプルとする.

2.2 分析方法

標本化周波数 12kHzで16ビットに量子化し、分析フレーム長30msec、分析周期５msecに設定して、ハミング窓を用い、$1-Z^{-1}$の特性で高域強調した後、1～16 次のＬＰＣケプストラム係数を抽出する. 各母音区間の中央フレームの特徴パラメータを入力層に入力する.

3. ニューラルネットの構造、教師信号、学習方法

ニューラルネットによる音素認識において、入力層に入力する特徴パラメータの違いによって認識性能に差が生じる[9]. これは、入力層

のユニットがはる空間での入力パターンの特徴ベクトルの分布と、出力層のユニットがはる空間での教師信号ベクトルの間のマッピングの自然さ（規則性の抽出しやすさ）の程度の違いによると考えられる.

本稿では、このことを、特徴パラメータを同じにして教師信号ベクトルへのマッピングを違えた場合について検討する.

3.1 ニューラルネットの構造

図1のような3層構造のニューラルネットを用いる. 入力層は16ユニット、中間層は9ユニットである. 出力層は1、2、5、16ユニットの4通りの場合について検討する. 中間層と出力層の各ユニットでは、入力値をシグモイド関数で変換したものが出力値となる.

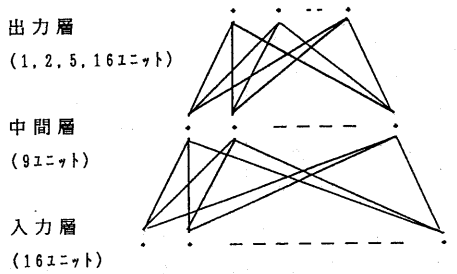


図1　3層構造のニューラルネット

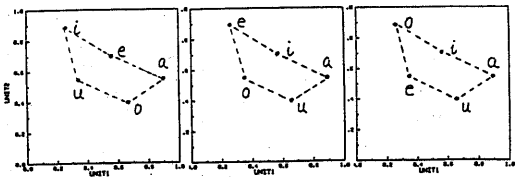
3.2 教師信号

(1) 出力ユニット数1の場合

0～1の間で等間隔に設定した5つの教師信号 0.1, 0.3, 0.5, 0.7, 0.9を、マッピング<1-A>ではいわゆる母音5角形の母音並びi,e,a,o,uの順に、マッピング<1-B>では単純にa,i,u,e,oの順に対応させる.

(2) 出力ユニット数2で、ホルマントの母音5角形を教師信号に用いる場合

0～1の間に正規化した第1、第2ホルマントの各母音の典型的な値 /a/=(0.9, 0.55)、/i/=(0.25, 0.9)、/u/=(0.35, 0.55)、/e/= (0.55, 0.7)、/o/=(0.65, 0.4)を教師信号として、マッピング<2-A>では図2(a)のようにそのままa,i,u,e,oの順に対応させる. マッピング<2-B>、<2-C>では 図2(b)(c)のようにその対応関係を故意に変える.



マッピング<2-A>　マッピング<2-B>　マッピング<2-C>

(a)　(b)　(c)

図2　ホルマントの母音5角形を教師信号に用いる場合の教師信号と母音の対応関係

(3) 出力ユニット数1,2,5,16で、平均ケプストラム係数を教師信号に用いる場合

特徴パラメータを学習サンプル内で母音毎に平均し、出力ユニット数1,2,5,16の場合に、それぞれ1次、1～2次、1～5次、1～16次の平均ケプストラム係数を教師信号として、そのままa,i,u,e,oの順に対応させるものをマッピング<1-a>、<2-a>、<5-a>、<16-a>とする. また、図2(b)(c)と同様に、その対応関係を故意に変えるものをマッピング<1-b>,<1-c>,･･･,<16-b>,<16-c>とする.

なお、この場合の出力層の各ユニットでは、シグモイド関数による変換を行わずに、入力値をそのまま出力値とする.

(4) 出力ユニット数5で、(1－ε、ε)の教師信号を用いる場合

入力のカテゴリ数と同数のユニットを出力層に並べて、正解ユニットの教師信号を1－ε、他のユニットの教師信号をεとする. εの値は0から0.2まで0.05きざみで変えて、それぞれマッピング<5-1.00>,<5-0.95>,<5-0.90>,<5-0.85>,<5-0.80>とする.

教師信号をまとめて表1に示す. この表の平均ケプストラム係数は 発声者katの音声資料から求めたものである.

3.3 学習方法

重み係数の学習は誤差逆伝搬法で行い、学習パラメータは、慣性率（momentum）$\alpha = 0.9$、学習率（learning rate）$\eta = 0.1$ とする. 学習は 10000回繰り返す.

学習サンプル、テストサンプル各々について、学習回数による認識率の立ち上がり方、学習中に得られる認識率の最大値、最大認識率を与え

表1　教師信号

タイプ	出力ユニット数		教師信号 /a/	/i/	/u/	/e/	/o/	マッピング
母5音角形	1	# 1	0.5	0.1	0.9	0.3	0.7	<1-A>
	2	# 1	0.9	0.25	0.35	0.55	0.65	<2-A>
		# 2	0.55	0.9	0.55	0.7	0.4	
平均ケプストラム係数	1	# 1	1.37	0.43	1.08	1.16	1.30	<1-a>
	2	# 2	0.09	0.06	0.31	-0.17	0.53	<2-a>
	5	# 3	-0.48	0.33	-0.42	-0.12	-0.10	
		# 4	0.26	0.95	0.50	0.69	0.27	<5-a>
		# 5	0.07	0.23	0.44	0.33	0.07	
	16	# 6	0.04	0.22	0.25	0.07	-0.21	
		# 7	-0.32	-0.11	-0.11	-0.33	-0.25	
		# 8	0.07	-0.04	-0.01	-0.09	-0.02	
		# 9	-0.08	0.00	0.17	-0.16	0.27	
		#10	-0.17	-0.02	-0.20	-0.16	-0.11	
		#11	-0.15	-0.02	-0.24	0.00	-0.17	<16-a>
		#12	0.01	-0.13	-0.08	-0.04	-0.09	
		#13	0.09	-0.13	-0.02	-0.10	-0.10	
		#14	0.08	-0.03	0.02	-0.11	-0.03	
		#15	0.00	-0.04	-0.10	-0.01	-0.08	
		#16	0.02	-0.10	-0.11	0.04	-0.07	
1-ε	5	# 1	1-ε	ε	ε	ε	ε	<5-(1-ε)>
		# 2	ε	1-ε	ε	ε	ε	
		# 3	ε	ε	1-ε	ε	ε	
		# 4	ε	ε	ε	1-ε	ε	
		# 5	ε	ε	ε	ε	1-ε	

る時の出力値ベクトルの分布、等を調べる．なお、認識率は、出力層のユニットがはる空間で、出力値ベクトルに（ユークリッド距離で）最も近い教師信号ベクトルに対応する母音を認識結果と判定して求める．

4．実験結果と考察

4.1 学習回数と認識率

マッピング <1-A>, <1-B>, <2-A>, <2-B>, <2-C> の各場合について、学習回数による認識率及び誤差の変化を図3 (a)～(e)に示す．○印は学習サンプルの認識率、×印はテストサンプルの認識率、△印は誤差を示す．図3 (a)～(e)の各々において、左側の図は学習回数1～400回における変化、右側の図は学習回数1～10000回における変化を示す．図中の矢印は学習サンプルあるいはテストサンプルの認識率が最大となるところである．

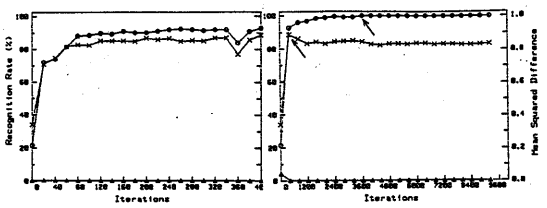
4.2 出力値ベクトルの分布

マッピング <1-A>, <1-B>, <2-A>, <2-B>, <2-C> の各場合について、学習サンプルあるいはテストサンプルの認識率が最大となる時（図3中の矢印で示されるところ）の出力値ベクトルの分布を図4 (a)～(e)に示す．図4 (a)～(e)の各々において、左側の図は学習サンプルの分布、右側の図はテストサンプルの分布を示す．マッピング <2-a>, <2-b>, <2-c>の各場合について、同様の出力値ベクトルの分布を図5 (a)～(c)に示す．図5 (d) には入力層に入る1～2次のケプストラム係数の分布を示す．

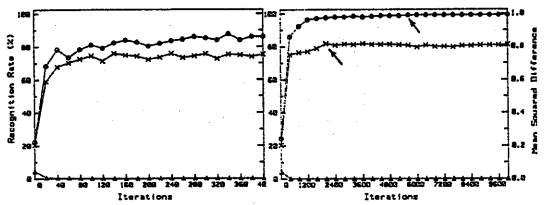
4.3 教師信号マッピングと認識率

学習中の最大認識率を、教師信号と母音の対応関係を決めるマッピングの各場合について、まとめて表2に示す．

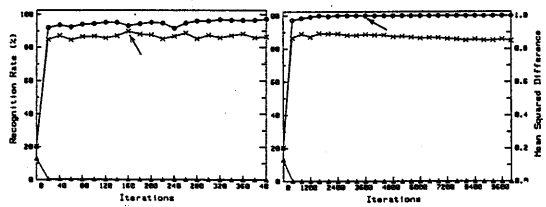
比較のために、表2の最下欄には、特徴パラメータを学習サンプル内で母音毎に平均したものを標準パターンとして、ニューラルネットを用いずに、入力音声の特徴ベクトルに（ユークリッド距離で）最も近い教師信号ベクトルに対応する母音を認識結果とした場合の認識率を示す．1次、1～2次、1～5次、1～16次の平均ケプ



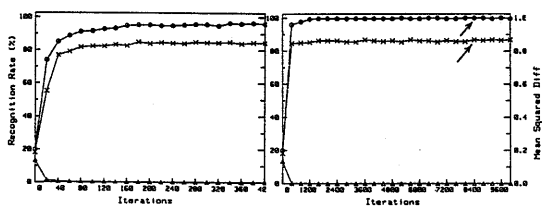
(a) マッピング〈1-A〉



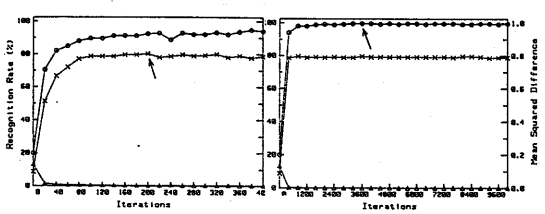
(b) マッピング〈1-B〉



(c) マッピング〈2-A〉

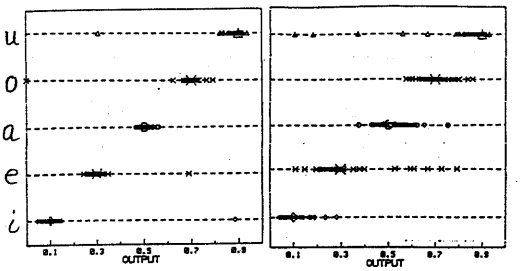


(d) マッピング〈2-B〉

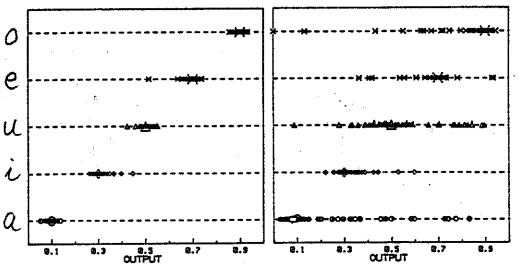


(e) マッピング〈2-C〉

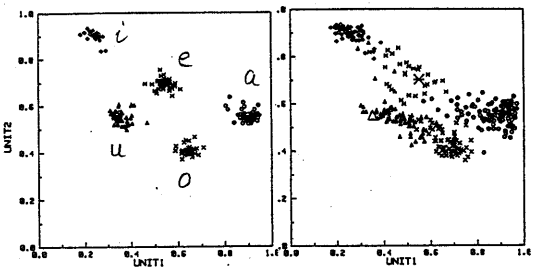
図3　学習回数による認識率及び誤差の変化
（発声者kat）



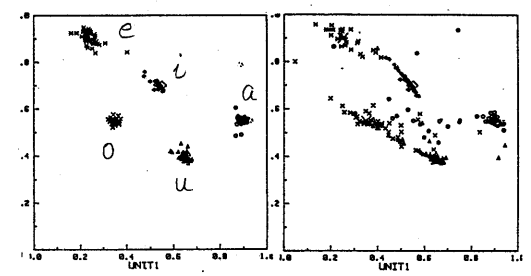
(a) マッピング〈1-A〉



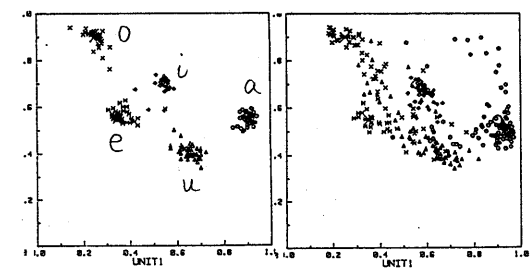
(b) マッピング〈1-B〉



(c) マッピング〈2-A〉

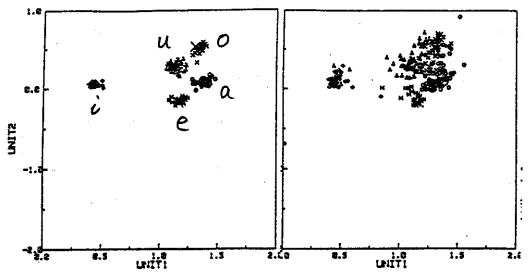


(d) マッピング〈2-B〉

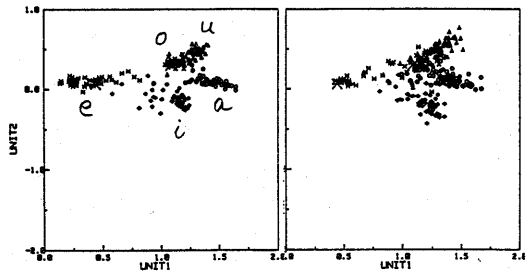


(e) マッピング〈2-C〉

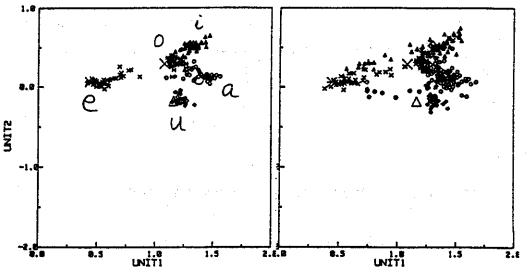
図4　出力値ベクトルの分布（発声者kat）



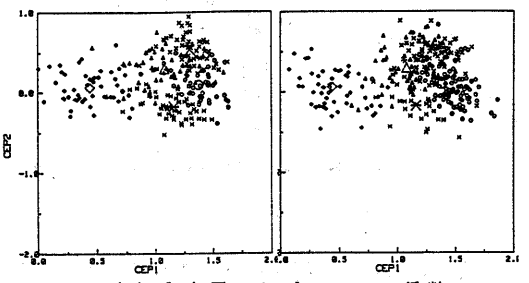
(a) マッピング<2-a>



(b) マッピング<2-b>



(c) マッピング<2-c>



(d) 入力層のケプストラム係数

図5　出力値ベクトルの分布　（発声者kat）

表2　認識率

マッピング		発声者 ito		発声者 kat	
		学習ﾃﾞｰﾀ	ﾃｽﾄﾃﾞｰﾀ	学習ﾃﾞｰﾀ	ﾃｽﾄﾃﾞｰﾀ
母音5角形	<1-A>	100	95.6	100	88.7
	<1-B>	100	91.2	99.2	81.8
	<2-A>	100	98.7	100	89.9
	<2-B>	100	97.2	100	87.0
	<2-C>	100	96.9	99.6	80.1
平均ケプストラム係数	<1-a>	93.2	85.6	94.0	67.0
	<1-b>	78.8	73.0	68.0	52.3
	<1-c>	80.8	81.2	71.6	61.6
	<2-a>	100	97.2	100	86.2
	<2-b>	96.0	91.3	92.4	84.3
	<2-c>	97.6	94.4	95.2	79.2
	<5-a>	100	95.0	99.6	86.5
	<5-b>	100	95.3	95.6	87.4
	<5-c>	99.6	94.3	96.0	84.3
	<16-a>	100	97.2	99.6	89.3
	<16-b>	99.2	95.6	96.0	86.5
	<16-c>	99.6	97.8	98.8	84.9
1-ε	<5-1.00>	99.6	97.5	98.8	88.7
	<5-0.95>	100	97.8	100	89.9
	<5-0.90>	99.6	97.5	100	89.9
	<5-0.85>	99.6	97.8	100	91.5
	<5-0.80>	99.6	97.8	100	89.9
標準パタン	<RP-1>	50.8	53.9	52.8	55.6
	<RP-2>	70.0	67.0	77.6	69.5
	<RP-5>	90.8	91.2	88.0	81.4
	<RP-16>	96.8	94.0	92.0	85.2

ストラム係数を標準パターンとして用いる場合を、それぞれ<RP-1>,<RP-2>,<RP-5>,<RP-16>で示した.

4.4 考察

（１）出力ユニット数と教師信号が同じであっても、教師信号と母音の対応関係を決めるマッピングの違いによって、学習の能率、出力値ベクトルの分布、認識の性能等に差がみられる.

- 図３の学習回数による認識率の変化をみると、学習サンプルでは、学習中の最大認識率はほとんど変わらないが、学習回数による認識率の立ち上がり方に差が生じる. テストサンプルでは、学習回数による認識率の立ち上がり方ばかりでなくて、学習中の最大認識率にも差が生じる.
- 図４の出力値ベクトルの分布をみると、学習サンプルでは、教師信号のまわりによくまとまって明確なクラスタを形成し、マッピングの違いによる差はみられない. テストサンプルでは、クラスタが崩れて分布が広がるときに、教師信号の間を埋めるように比較的自然な形で散らばる場合と、一部のサンプルが突飛な散らばり方をする場合がある.
- 表２の認識率をみると、マッピング<1-A>,<1-B>の比較や、マッピング <2-A>,<2-B>,<2-C>の比較からわかるように、教師信号が同じであってもマッピングの違いによってテストサンプルの認識率に差が生じる.

（２）出力ユニット数１の場合はマッピング<1-A>、出力ユニット数２の場合はマッピング<2-A>の認識率が良い. ホルマントの母音５角形に基づいて教師信号を設定し、母音スペクトル間の距離の相対的な関係をできるだけ反映したマッピングは、より少ない出力ユニット数で、他の教師信号を用いたマッピング<5-(1-ε)>やマッピング<16-a>と同等の認識率が得られている.

一般に、入力層のユニットがはる空間での入力パターンの特徴ベクトルの分布と、出力層のユニットがはる空間での教師信号ベクトルの間のマッピングがより自然となるように教師信号ベクトルを割り当てると認識率が高くなる.

なお、マッピング <2-A>は一種のホルマント抽出とみることもできるが、ニューラルネットの出力値と音声サンプルの実際のホルマントとの対応の分析が必要である.

（３）正解ユニットの教師信号を１－ε、他のユニットの教師信号をεとするマッピングでは、正解ユニットの教師信号が0.85のとき認識率が良い. これは文献[8]と同様の結果である. 各母音の教師信号ベクトルはεの値によらず互いに等距離にあるので、この場合の認識率の違いは、マッピングの自然さの程度よりむしろ、シグモイド関数の非線形性に対する要求の度合を反映していると考えられる[8].

（４）平均ケプストラム係数を標準パターンとして用いる場合の認識率は、その次数と同じ出力ユニット数のニューラルネットによる認識率より低い. これは、ニューラルネットを用いて非線形なマッピングを行うこと自体の有用性を示し、次数が小さいときほどその効果が大きい.

5. むすび

多層パーセプトロン型のニューラルネットによる母音認識の学習において、入力音声の性質を考慮した教師信号設定法を検討した. その結果、ホルマントの母音５角形に基づいて教師信号を設定し、母音スペクトル間の距離の相対的な関係をできるだけ反映したマッピングは、より少ない出力ユニット数で、他の教師信号を用いる場合と同等の認識率が得られることを示した.

このことから、一般に、入力層のユニットがはる空間での入力パターンの特徴ベクトルの分布と、出力層のユニットがはる空間での教師信号ベクトルの間のマッピングがより自然となるように、言い換えれば、入力層と出力層の間のマッピングの規則性が抽出しやすいように、カテゴリ間の距離の相対的な関係に関する事前の知識を考慮して教師信号を設定することが、ニューラルネットの学習を能率良く行う上で重要であるといえよう.

文　献

[1] “ニューロコンピューティング論文特集”、信学論（D-II）、J73-D-II、8、pp.1101-1359（平2-08）.

[2] 中川、鹿野、東倉：“音声・聴覚と神経回

路網モデル”、オーム社、6章 人工ニューラルネットワークによる音声情報処理、pp.185-224（1990）.
[3] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang : “ Phoneme recognition using time-delay neural networks ”, IEEE Trans. ASSP, ASSP-37, 3, pp.328-339（March 1989）.
[4] H. Sakoe, R. Isotani, K. Toshida, K. Iso, T. Watanabe : “ Speaker-independent word recognition using dynamic programming neural networks”, ICASSP89, S1.8, pp.29-32（May 1989）.
[5] 磯、渡辺：“ニューラル予測モデルを用いた不特定話者音声認識”、信学論（D-Ⅱ）、J73-D-Ⅱ、8、pp.1315-1321（平2-08）.
[6] J. Tebelskis, A. waibel : “ Large vocabulary recognition using linked predictive neural networks ”, ICASSP90, S8.7, pp.437-440（April 1990）.
[7] H. Leung, V. Zue : “ Some phonetic recognition experiments using artificial neural nets”, ICASSP88, S10.4, pp.422-425（April 1988）.
[8] 入野、河原：“多層神経回路網の非線形多変量解析による構成法－不特定話者母音認識への適用－”、信学論（D-Ⅱ）、J72-D-Ⅱ、8、pp.1187-1193（平1-08）.
[9] 中村、鹿野：“時間遅れ神経回路網（TDNN）における入力パラメータの評価と話者適応化”、信学技報、SP89-18（平1-06）.