



HOKKAIDO UNIVERSITY

Title	認識回路網の最適化による音声認識
Author(s)	吉岡, 浩; 青木, 由直
Citation	北海道大學工學部研究報告, 84, 83-89
Issue Date	1977-07-11
Doc URL	https://hdl.handle.net/2115/41411
Type	departmental bulletin paper
File Information	84_83-90.pdf



認識回路網の最適化による音声認識

吉 岡 浩 青 木 由 直

(昭和 51 年 12 月 28 日受理)

Speech Recognition by Optimization of Recognition Networks

Hiroshi YOSHIOKA and Yoshinao AOKI

(Received December 28, 1976)

Abstract

In this paper, we discuss optimization of recognition networks and have made an experiment to apply the technique to the recognition of speech.

A recognition network consists of the standard elements that have weighting resistors and a nonlinear logic circuit. The optimization of recognition network is achieved under the condition that a sum of false alarm and miss rates have a minimum value. The minimum point is found by a simplex method that is used to minimize a function of multi-variables. The advantages of this technique are the one-to-one correspondence between the optimization process and physical realization of the recognition network, and the real-time recognition. Whereas the disadvantage is that the optimization by this technique generally terminates on a local minimum.

Speech data are sampled from a continuous speech, a weather forecast broadcasted by NHK. As the recognition parameter we use coefficients of the digital inverse filter obtained through the short-term analysis.

1. ま え が き

荷重抵抗と非線形論理回路から成る認識回路網であっても効果的な学習によって高精度の認識が可能となる¹⁾。この認識回路網の判別関数は対象の分布関数を推定することなしに決定できる。ここでは認識の対象を日本語の 5 母音とし、それぞれのパターンを分離するのに適当と思われる認識回路網を仮定した。認識回路網には音声の短区間分析から得られるディジタル逆フィルタの係数を規格化して与え、シンプレックス法によって最適値を求めた。この認識方法には次のような特徴がある。認識回路網の構成、最適値を捜す過程と物理的実現との間には対応が付き、ハードウェアによって構成することができる。また認識はほとんど実時間で行なえる。一方、シンプレックス法は一般に極大値に収束するために始点の設定によっては最大値とならないことがある。

認識に用いたデータはラジオの気象通報の中の地名であり、山本らの製作によるマイクロコンピュータによる音声分析装置²⁾によって、区間再生しながら適当な区間を収集した。

2. 認 識 回 路 網

単位識別要素としては図 1 (a) に示す荷重抵抗と非線形回路から成る回路網を用いる。これは

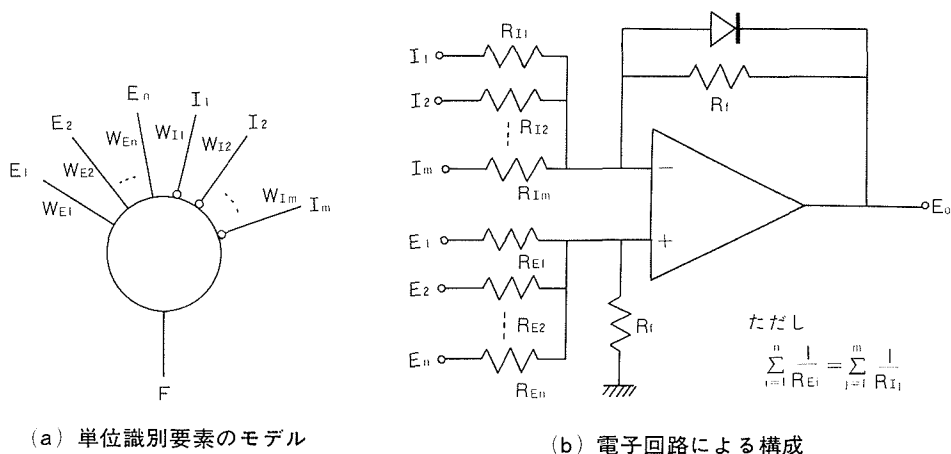


図1 単位識別要素

過渡応答を考慮しない、単純化したニューロンモデルでもある。この単位識別要素の入出力関係を次のように定める。

$$F = \begin{cases} 0 & W_E E - W_I I < \theta_2 \text{ のとき} \\ W_E E - W_I I & \theta_1 \leq W_E E - W_I I \leq \theta_2 \text{ のとき} \\ \theta_2 & W_E E - W_I I > \theta_2 \text{ のとき} \end{cases} \quad (1)$$

ここで

$$\begin{aligned} E &= (E_1, E_2, \dots, E_n)^T \\ I &= (I_1, I_2, \dots, I_m)^T \\ W_E &= (w_{E1}, w_{E2}, \dots, w_{En}) \\ W_I &= (w_{I1}, w_{I2}, \dots, w_{Im}) \end{aligned}$$

E は興奮性入力, I は抑制性入力と呼ばれ, W_E, W_I は荷重抵抗, θ_1, θ_2 は閾値である。

上述の単位識別要素は図1 (b) に示す演算増幅器を用いた電子回路によって容易に実現することができる。この回路を用いれば演算出力は瞬時に得られる。出力は次のようになる。

$$E_0 = R_f \left\{ \frac{\left(1 + R_f \sum_{j=1}^m \frac{1}{R_{Ij}}\right)}{\left(1 + R_f \sum_{i=1}^n \frac{1}{R_{Ei}}\right)} \sum_{i=1}^n \frac{E_i}{R_{Ei}} - \sum_{j=1}^m \frac{I_j}{R_{Ij}} \right\} \quad 0 \leq E_0 \leq V_0 \quad (2)$$

ただし V_0 は演算増幅器の電源電圧である。

3. 認識回路網の最適化方法

前章で述べた単位識別要素を適当に組み合わせて特定の音声にだけ反応する認識回路網を構成する。この回路網の荷重抵抗 $W = (w_1, w_2, \dots, w_N)$ は教育用データを与えた時に回路網の認識率が最大となるように定めるものとする。これは荷重抵抗 W を座標点とし、認識率をその点の関数値とする多次元空間における最大最小問題である。ところが与えられた関数が多峰性の場合にはその最大最小を求める有効な手法は今だ開発されていない。そこでここでは単峰性関数の最大最小問題に効果的であるシンプレックス方式を用いることにする。音声認識回路網の認識率は荷重抵抗 W に対して多峰性関数になっていることが予想される。しかし荷重抵抗の初期値はある程度予測がつき、探索範囲を最大値の近傍の領域に限定すれば単峰性とみなせると思う。

シンプレックス方式⁹⁾は N 次元空間内の $(N+1)$ 点のうちで関数値が最小の点を新しい点に転位させながら全体を徐々に最大値のまわりに収束させていくものである。

(i) N 次元空間で始点 $w_{1k} = (w_{11}, w_{12}, \dots, w_{1N})$ が与えられた時, 残りの N 点, $w_{2k}, w_{3k}, \dots, w_{N+1,k}$ は次式で与えられる。

$$w_{jk} = \begin{cases} w_{1k} + p & j = k+1 \text{ のとき} \\ w_{1k} + q & j \neq k+1 \text{ のとき} \end{cases} \quad \begin{matrix} \text{ただし } j = 2, 3, \dots, N+1 \\ k = 1, 2, \dots, N \end{matrix} \quad (3)$$

ここで

$$p = \frac{N-1+\sqrt{N+1}}{\sqrt{N+1}-1} \quad q = 1$$

$w_{1k}, w_{2k}, \dots, w_{N+1,k}$ は N 次元空間内で正 $(N+1)$ 面体の頂点となっている。

(ii) 認識率は $f(\alpha, \beta) = 1 - \alpha - \beta$ で定義する。

ここで

$$\alpha = \frac{\text{受け入れるべきものの中で誤まった数}}{\text{受け入れるべきパターン数}}$$

$$\beta = \frac{\text{受け入れるべきでないものの中で誤まった数}}{\text{受け入れるべきでないパターン数}}$$

α エラー率, β エラー率に重みを乗じないでいるのは次の理由による。一般に認識系では α エラーの方が β エラーより損失が大であると考えられる。そのため α エラーに重み付けしてその比重を大きくした方が妥当と思われるが, 多クラス分類機械では α エラーの定義式の分母が β エラーのそれより小さくなるため, 単一エラーがより大きな損失を持つことになる。そのため重み付けは行なわなかった。

(iii) 計算機で次のものを計算する。

$$W * M = \begin{pmatrix} \alpha_1 & \beta_1 & f(\alpha_1, \beta_1) \\ \alpha_2 & \beta_2 & f(\alpha_2, \beta_2) \\ \vdots & \vdots & \vdots \\ \alpha_{N+1} & \beta_{N+1} & f(\alpha_{N+1}, \beta_{N+1}) \end{pmatrix} \quad \text{ここで} \quad W = \begin{pmatrix} w_{11} & w_{12} \dots & w_{1N} \\ w_{21} & w_{22} \dots & w_{2N} \\ \vdots & \vdots & \vdots \\ w_{N+1,1} & w_{N+1,2} \dots & w_{N+1,N} \end{pmatrix} \quad M = \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_N \end{pmatrix} \quad (4)$$

M は入力データの組であり, $*$ は認識回路網に入力データ M が加わった時, $\alpha_j, \beta_j, f(\alpha_j, \beta_j)$ が生ずることを示す。

(iv) シンプレックス方式で最低の点を新しい点に移して行くには次のように行なう。

① $f_i = \min\{f(\alpha_j, \beta_j)\}$ を見つける。

② 反射 w_{lk} を w_{lk} を除いた点の重心に関して対称な位置に移し, その点の値を f_i^* とする。

$$w_{lk}^* = \frac{2}{N} \sum_{j=1}^{N+1} w_{jk} - \left(1 + \frac{2}{N}\right) w_{lk} \quad (k=1, 2, \dots, N)$$

③ 拡張 $f_i^* \geq f_i$ のときは w_{lk}^* の点を w_{lk} と w_{lk}^* を結んだ延長上にとる。

$$w_{lk}^{**} = \frac{3}{N} \sum_{j=1}^{N+1} w_{jk} - \left(2 + \frac{3}{N}\right) w_{lk} \quad (k=1, 2, \dots, N)$$

$f_i^{**} < f_i^*$ のときは w_{lk}^* を新しい位置として①に戻る。

④ 縮小 $f_i^{**} < f_i$ のときは w_{lk}^{**} の点を w_{lk} と w_{lk} を除いた点の重心との中心にとり, その点での f_i^{***} を求める。 $f_i^{***} > f_i$ のときは w_{lk}^{***} を新しい点として①に戻る。

⑤ 圧縮 $f_i^{***} \leq f_i$ のときはすべての点を $f_h = \max\{f(\alpha_j, \beta_j)\}$ なる点 w_{hk} のまわりに圧縮する。

$$w_{jk}^{****} = \frac{w_{jk} + w_{hk}}{2}$$

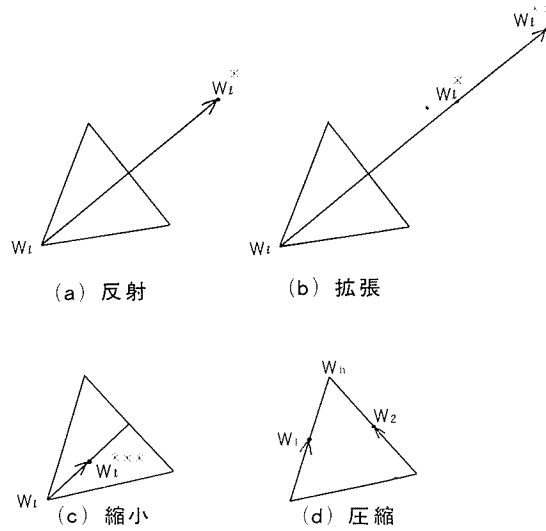
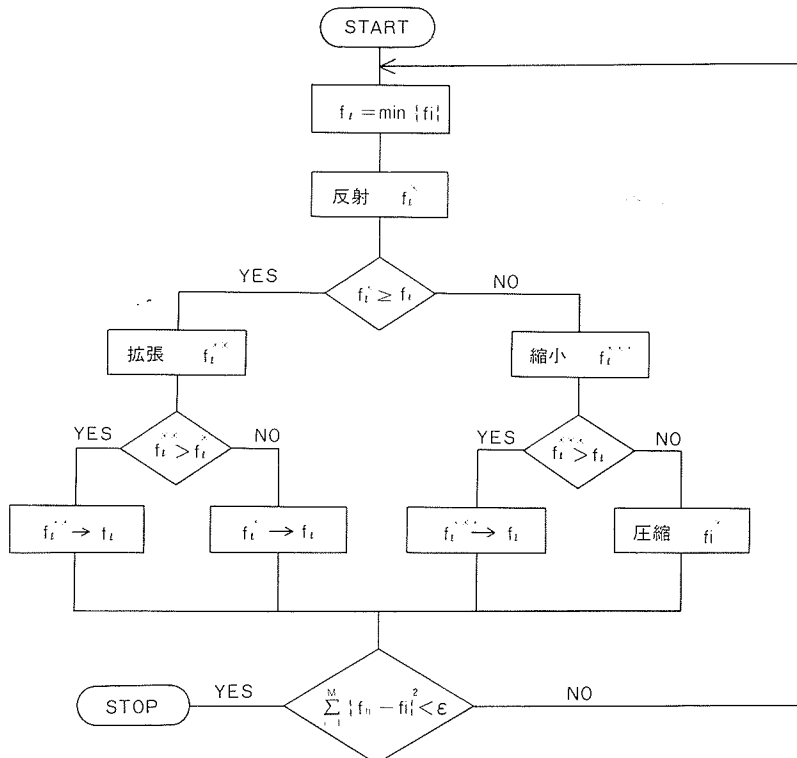
図2 シンプレックス法による最適化の手順 ($N=2$)

図3 シンプレックス法のフローチャート

図2は上に述べた反射，拡張，縮小，圧縮の過程を $N=2$ の場合を例にとって図示したものである。この過程は式 (5) が成立した時に停止するものとする。 ϵ としては 10^{-4} を与えた。

$$\sum_{j=1}^{N+1} (f_h - f_j)^2 < \epsilon \quad (5)$$

図3はシンプレックス方式のフローチャートである。

4. 認識パラメータの決定

第2章で述べた認識回路網を音声認識に適用する。その場合の認識パラメータとしてはデジタル逆フィルタの係数⁴⁾を用いることにする。ここでは逆フィルタの推定方法について簡単に述べる。

4.1 音声生成モデルと逆フィルタ

音声信号は一般には定常でないが、適当な短区間内では定常と考えてよい。また母音のみを考えればそれは零点を含まず、極だけで表現することができる。以上を考慮すれば音声の生成モデルは図4で表わすことができ、その伝達関数 $G(z)$ は式 (6) となる。

$$G(z) = \sigma \cdot \frac{1}{1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \dots + \alpha_n z^{-n}} \quad (6)$$

音声生成モデルの逆フィルタは $G(z) = \sigma \cdot 1/H(z)$ なる関係があるから式 (7) となる。

$$H(z) = 1 + \sum_{i=1}^n \alpha_i z^{-i} \quad (7)$$

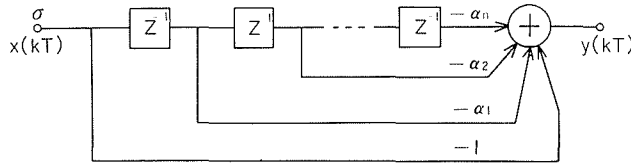


図4 音声生成モデル

4.2 逆フィルタの推定

逆フィルタの入力として音声信号を加えた時に出力に現われるのは声帯の励振波形である。声帯振動波は個人によって大きく異なり、また同一発声者でも強さや音の高さによって変化するが、ここでは簡単のため単一インパルスと仮定する。逆フィルタが安定であればその出力 y_k は

$$y_k = x_k + \sum_{i=1}^p \alpha_i x_{k-i} + e_k \quad (8)$$

但し、出力 y_k は単一インパルスと仮定したから y_k は $k=0$ のときのみ1で $k \neq 0$ のときは0である。また e_k は残差のサンプル値である。式 (8) は行列の形で式 (9) とも書ける。

$$Y = X_0 + X\alpha + E \quad (9)$$

式 (9) より α の推定値 $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_p)^T$ を最小二乗推定法で求める。 $\hat{\alpha}$ は式 (10) を解くことによって求まる。

$$X^T(Y - X_0) = X^T X \hat{\alpha} \quad (10)$$

Y は逆フィルタの出力のサンプル値を要素とする行列で $k=0$ のときのサンプル値を含んでいないから式 (10) は次のようになる。

$$-X^T X_0 = X^T X \hat{\alpha} \quad (11)$$

音声信号に式 (11) を適用して $\hat{\alpha}$ を推定する場合、観測波形の N 個だけのサンプル値を用いる短区間分析を行なう。そのため $i \leq 0$, $i > N$ では $x_i = 0$ であるから式 (11) は式 (12) のようになる。

$$-\begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{pmatrix} = \begin{pmatrix} R_0 & R_1 & \dots & R_{p-1} \\ R_1 & R_0 & \dots & R_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & \dots & R_0 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_p \end{pmatrix} \quad \begin{matrix} \text{ただし} \\ R_i = \sum_{k=1}^{N-i} x_k x_{k+i} \\ i = 0, 1, \dots, p \end{matrix} \quad (12)$$

式 (12) の連立一次方程式を解くことによって $\hat{\alpha}$ を求めることができる。

5. 母音認識への適用

第 2 章で述べた単位識別要素を適当に組み合わせた認識回路網に、音声の短区間分析で得られた規格化したデジタル逆フィルタの係数を与え、回路網が最適化していく過程、また最適化し

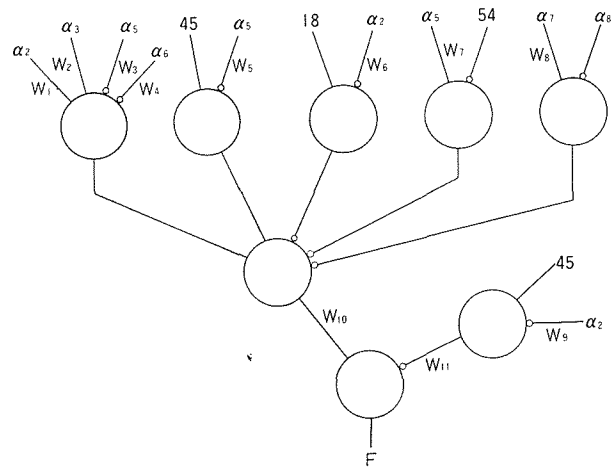


図 5 / オ / に対する認識回路網

表 1 / オ / に対する最適化の過程

シンプ ックス番号	荷 重 抵 抗		w_1, w_2, w_3, w_4 w_5, w_6, w_7, w_8 w_9, w_{10}, w_{11}	α	β	$1-\alpha-\beta$	繰り返 し 数
1	0.9	0.9	0.9 0.9	0.0833	0.1712	0.7455	4
9 4 (i) 11 (ii) 3 (iii)	0.9	0.9	0.9 0.9				
	0.9	0.9	0.9				
	0.7406	0.9133	0.9254 2.0201	0.0833	0.0721	0.8446	
	0.8227	0.7958	0.7453 1.0949	0.0833	0.0631	0.8536	
	1.4039	0.7509	0.9848	0.0417	0.0901	0.8682	5
				0.0417	0.0811	0.8773	22

(i) 反射, (ii) 拡張, (iii) 圧縮

表 2 5 母音に対する認識結果

入力数	ア	イ	ウ	エ	オ	子 音	α	β	$1-\alpha-\beta$
	33	24	19	21	24	14			
ア	33	0	0	1	0	0	0.0	0.0098	0.9902
イ	0	23	5	4	9	11	0.0417	0.2613	0.6971
ウ	0	6	17	0	2	7	0.1053	0.1293	0.7654
エ	0	7	1	20	1	2	0.0476	0.0955	0.8559
オ	0	0	3	0	22	3	0.0833	0.0541	0.8626

た時にどの程度の認識率が得られるかを実験した。実験に用いた音声データは NHK の気象通報の中の地名である。話者は男性アナウンサー 2 名である。放送はカセットテープに録音された後、山本らによって製作されたマイクロコンピュータによる音声分析装置によって 8 KHz のサンプリング周波数でサンプリングされた。量子化は 8 ビットである。このデータの中から区間再生を行ないながら 5 母音 / ア, イ, ウ, エ, オ / と若干の有声子音 / N, M, D, W, R, H / を収集した。ディジタル逆フィルタは 16 m sec, 128 点を 1 区間としたものにハミング窓を乗じて求めた。

シンプレックス法によって認識回路網が最適化していく様子を母音の / オ / の認識過程を例にとって示す。図 5 は母音の / オ / のときだけ出力 $F > 0$ となり、/ オ / 以外ときは $F = 0$ となるような認識回路網である。荷重抵抗は 11 個であるからシンプレックスの頂点の数は 12 となる。表 1 の第 1 段は始点の荷重抵抗およびその時の認識率である。第 2 段は 12 のシンプレックスの頂点のうちで最高の認識率のものである。以下は過去の最高点を越えた時のシンプレックスの頂点とそれが反射, 拡張, 縮小, 圧縮のうちでどの操作のときに越えたか, またそれまでに要した検索の繰り返し数を示している。

表 2 には日本語の 5 母音に対し、最適化された後の認識回路網による認識結果を示す。この認識実験では排他性を仮定していないので認識されたものの総和は一般には入力数よりも大きくなる。得られた認識結果は従来までに報告されているものほど良くない。その理由としては 1) 母音の認識に対し、子音のデータも加えている。2) 逆フィルタの係数は認識パラメータとしては不適当である。3) 認識回路網の設定が不完全である。などが考えられる。

6. あ と が き

荷重抵抗と非線形論理回路から成る認識回路網に連続音声の短区間分析で得られたディジタル逆フィルタの係数を教育用データとして用い、シンプレックス法により回路網を最適化する実験を行なった。認識回路網は試行錯誤的に組み立てねばならず、得られた母音の認識結果も従来の方法ほど良くないが、認識は実時間で行なえる利点がある。また物理的対応がつくため、ハードウェアで容易に実現することができる。

参 考 文 献

- 1) Harris, D. R.: Computer Optimization of Recognition Networks. IEEE Transaction on Computers vol, C-13, No. 10, p. 918-923 (Oct. 1969).
- 2) 山本 強, 青木由直: 工学部研究報告, 第 28 号, p. 59-65 (昭和 51 年 12 月).
- 3) J. A. Nelder and R. Mead: A Simplex Method for Function Minimization; Computer Journal p. 308-313, (Jan, 1965).
- 4) John D. Markel: Digital Inverse Filtering-A New Tool for Formant Trajectory Estimation; IEEE Trans. AU-20, p. 129-137 (Jun. 1972).