Internal Use Only

002

TR-IT-0179

重回帰写像モデルに基づく話者適応方式の検討

A Study on Speaker Adaptation Using Linear Regression

藤岡 慎治　　石井 純
Shinji fujioka　　Jun Ishii

1996.8.30

連続分布型 HMM を用いた話者適応の一つとして重回帰写像モデルに基づく方式があり、その有効性が報告されている。本稿では、最尤を基準に変換行列を推定する Maximum Likelihood Linear Regression (MLLR) について検討する。この方式では、適応データの量に応じて変換パラメータ数を最適に設定することが重要である。本稿では、適応データ数と変換パラメータ数の認識性能に与える影響を評価する。また、適応データ数により変換パラメータ数を決定する方法を検討し、音素認識実験で評価した結果を報告する。

ⓒＡＴＲ音声翻訳通信研究所

©ATR Interpreting Telecommunications Research Laboratories

もくじ

1 はじめに **1**

2 重回帰モデルに基づいた話者適応方式 **2**

1 重回帰写像モデルを用いた話者適応 2

1.1 モデルパラメータの適応 2

2 変換行列の推定 3

2.1 Baum の補助関数の導入 3

2.2 Baum の補助関数の最大化による変換行列の推定 5

3 変換行列の種類 5

4 データ量によるパラメータ数制御方式 6

3 評価実験 **8**

1 実験条件 8

2 実験結果 9

2.1 適応データ数による評価結果 11

2.2 変換パラメータ数による評価結果 16

2.3 データ量によるパラメータ数制御方式の評価結果 20

2.4 他の話者適応方式との比較 21

4 結論 **22**

参考文献 24

A 音素認識実験実験結果 **25**

1 変換パラメータ数による評価結果 25

2 音素認識実験結果 (各話者) 28

第 1 章

はじめに

音声認識技術の進歩により、連続分布型 HMM による不特定話者音声認識システムは高い性能を得られるようになった。しかし、特定話者音声認識システムと比較すると十分な能力が得られているとは言えない。不特定話者認識の性能を向上する方法として、不特定話者用音響モデルを入力話者の音声により話者適応する方法がある。

連続分布型 HMM を用いた話者適応の一つとして重回帰写像モデルに基づく方式 [1][2][3] がある。重回帰写像モデルに基づく話者適応では、不特定話者システムのモデルパラメータは重回帰に基づく平均ベクトルの変換を用いて入力話者のものに変換される。この話者適応法では、結びにより回帰変換行列の共有化を行なえるため、少量の適応データで話者適応を行なうときの未学習モデルと推定誤差の問題に対してロバストな適応が行なえる。本稿では、最尤を基準に変換行列を推定する Maximum Likelihood Linear Regression(MLLR)[1] について検討する。

MLLR を用いた話者適応の問題点として認識性能が適応データ量に依存することが挙げられる。認識性能は回帰変換行列のパラメータ数と共有化におけるクラス数により変化し、最適な回帰変換行列のパラメータ数と共有化におけるクラス数は適応データ量により変化する。このため、この方式では適応データの量に応じて変換パラメータ数を最適に設定することが重要である。

本稿では、MLLR を用いた話者適応においてデータ量に依存せずに最適な話者適応性能を得ることを目的とし、適応データ量と変換パラメータ数の認識性能に与える影響を評価し、適応データ量に応じて最適な回帰変換行列のパラメータ数と共有化におけるクラス数を選択する方法を検討する。

第 2 章

重回帰モデルに基づいた話者適応方式

1　重回帰写像モデルを用いた話者適応

1.1　モデルパラメータの適応

重回帰写像モデルによる適応は、初期モデルのガウス分布の平均ベクトル μ_k(k はガウス分布の番号) を拡張したベクトル ξ_k を用いて、変換行列 W によって、適応平均ベクトル $\hat{\mu}_k$ に変換することで行なわれる。

$$\hat{\mu}_k = W\xi_k \tag{2.1}$$

$$\xi_k = [1, \mu_{k_1}, \mu_{k_2}, \cdots, \mu_{k_n}]^T$$

ここで n は平均ベクトルの次元数、W は $n \times (n+1)$ の変換行列であり、適応データによって推定される。適応処理のブロック図を図 2.1に示す。

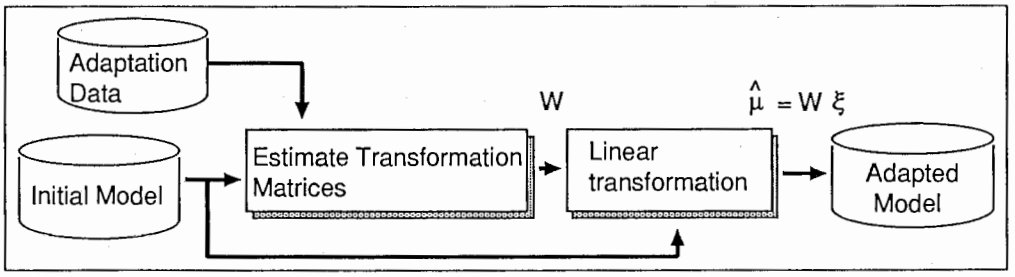

図 2.1: 重回帰写像モデルによる話者適応のブロック図

ここでは、適応後のモデルにおけるガウス分布 k の出力確立密度関数を式 (2.2) と仮定し、尤度最大を基準に変換行列 W を推定する Maximum Likelihood Linear Regression (MLLR) [1] を用いる。

$$b_k(\mathbf{o}_t) = \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}_t - W\xi_k)^T\Sigma_k^{-1}(\mathbf{o}_t - W\xi_k)\right] \tag{2.2}$$

ここで、$\mathbf{o}_t$ は時間 t における観測ベクトル、Σ_k はガウス分布 k の対角共分散行列 $diag(\sigma_{k_1}^2, \ldots, \sigma_{k_n}^2)$ を表している。

2 変換行列の推定

図 2.2のように、変換行列 W がガウス分布の集合 K で共有化されている場合の変換行列の推定について述べる。

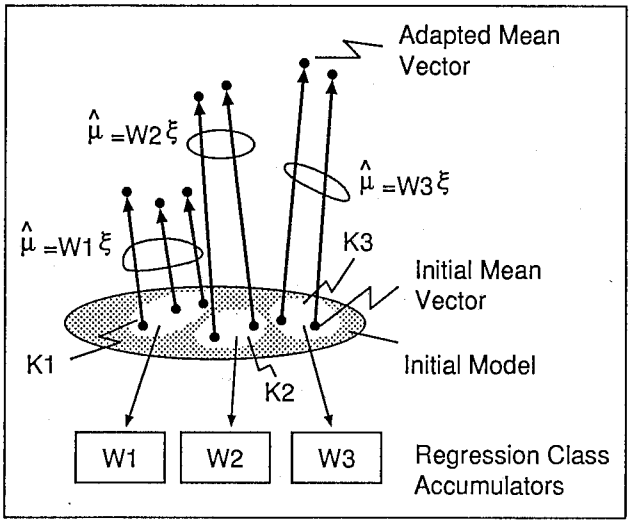

図 2.2: 変換行列の共有化

2.1 Baum の補助関数の導入

変換行列の推定は式 (2.3) に示されている Baum の補助関数の最大化によって行なわれる。

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} \mathcal{F}(O, \theta|\lambda) \log(\mathcal{F}(O, \theta|\bar{\lambda})) \tag{2.3}$$

ここで、O は適応データの特徴ベクトルの系列

$$\mathbf{o}_1, \ldots, \mathbf{o}_T \tag{2.4}$$

を、θ は状態の系列

$$\theta = (\theta_0, \theta_1, \ldots, \theta_T) \tag{2.5}$$

を表している。λ、$\bar{\lambda}$ は適応前後のモデルパラメータであり、$\mathcal{F}(O, \theta|\lambda)$、$\mathcal{F}(O, \theta|\bar{\lambda})$ は状態系列 θ に対する適応前後の尤度である。$a_{i,j}$ を状態 i から j への遷移確立、、K_{θ_t} を時間 t における状態の混合成分の集合、c_k を混合成分の重みとすれば、状態系列 θ に関する尤度は下

式で表される。

$$\mathcal{F}(O,\theta|\lambda) = a_{\theta_T,N}\prod_{t=1}^{T} a_{\theta_{t-1},\theta_t}\left[\sum_{k\in K_{\theta_t}} c_k b_k(\mathbf{o}_t)\right] \tag{2.6}$$

$$= \sum_{k\in K_{\theta_1}}\cdots\sum_{k\in K_{\theta_T}}\left[a_{\theta_T,N}\prod_{t=1}^{T} a_{\theta_{t-1},\theta_t} c_k b_k(\mathbf{o}_t)\right] \tag{2.7}$$

従って、補助関数は

$$Q(\lambda,\bar{\lambda}) = \sum_{\theta\in\Theta}\mathcal{F}(O,\theta|\lambda)\log\mathcal{F}(O,\theta|\bar{\lambda}) \tag{2.8}$$

$$Q(\lambda,\bar{\lambda}) = \sum_{\theta\in\Theta}\mathcal{F}(O,\theta|\lambda)\sum_{k\in K_{\theta_1}}\cdots\sum_{k\in K_{\theta_T}} f(\bar{a},\bar{b},\bar{c}) \tag{2.9}$$

$$f(\bar{a},\bar{b},\bar{c}) = \left[\log\bar{a}_{\theta_T N} + \sum_{t=1}^{T}\log\bar{a}_{\theta_{t-1}\theta_t} + \sum_{t=1}^{T}\log\bar{b}_k(\mathbf{o}_t) + \sum_{t=1}^{T}\log\bar{c}_k\right] \tag{2.10}$$

と再定義され、変換行列の推定は出力確立にのみ依存するので

$$Q(\lambda,\bar{\lambda}) = constant + \sum_{\theta\in\Theta}\mathcal{F}(O,\theta|\lambda)\sum_{k\in K_{\theta_1}}\cdots\sum_{k\in K_{\theta_T}}\sum_{t=1}^{T}\log\bar{b}_k(\mathbf{o}_t) \tag{2.11}$$

と簡単化できる。

さらに、共有化されたガウス分布の集合 K における補助関数を Q_K は下でで定義される式となる。

$$Q_K(\lambda,\bar{\lambda}) = constant + \sum_{k\in K}\sum_{t=1}^{T}\mathcal{F}(O,k_t=k|\lambda)\log\bar{b}_k(\mathbf{o}_t) \tag{2.12}$$

今、$\gamma_k(t)$ を、時刻 t にガウス分布 k でデータが観測される期待値

$$\gamma_k(t) = \frac{1}{\mathcal{F}(O|\lambda)}\mathcal{F}(O,k_t=k|\lambda) \tag{2.13}$$

であるとすると式 (2.12) は

$$Q_k(\lambda,\bar{\lambda}) = constant + \mathcal{F}(O|\lambda)\sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)\log b_k(\mathbf{o}) \tag{2.14}$$

さらに、式 (2.2) で定義された出力確立密度関数により、

$$Q_K(\lambda,\bar{\lambda}) = constant - \frac{1}{2}\mathcal{F}(O|\lambda)\sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)[n\log(2\pi) + \log|\Sigma_k| + h(\mathbf{o}_t,k)] \tag{2.15}$$

$$h(\mathbf{o}_t,k) = (\mathbf{o}_t - W\xi_k)^T\Sigma_k^{-1}(\mathbf{o}_t - W\xi_k) \tag{2.16}$$

に書き換えられる。

2.2 Baumの補助関数の最大化による変換行列の推定

ガウス分布の集合 K の共有化において、補助関数の最大を示す変換行列は、補助関数 Q_K を W に関して偏微分を行ない、零とした

$$\frac{dQ_K(\lambda,\bar{\lambda})}{dW} = \mathcal{F}(O|\lambda)\sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)\Sigma_k^{-1}[\mathbf{o}_t - W\xi_k]\xi_k^T = 0 \tag{2.17}$$

$$\sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)\Sigma_k^{-1}\mathbf{o}_t\xi_k^T = \sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)\Sigma_k^{-1}W\xi_k\xi_k^T \tag{2.18}$$

によって算出される。従って式 (2.18) より、変換行列 W の p 行目 $w_{p,i}$ は下式によって与えられる値となる。

$$\begin{pmatrix} w_{p,1} \\ \vdots \\ w_{p,n+1} \end{pmatrix} = \begin{pmatrix} g_{1,1}^{(p)} & g_{1,2}^{(p)} & \cdots & g_{1,n+1}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n+1,1}^{(p)} & g_{n+1,2}^{(p)} & \cdots & g_{n+1,n+1}^{(p)} \end{pmatrix}^{-1} \begin{pmatrix} z_1^{(p)} \\ \vdots \\ z_{n+1}^{(p)} \end{pmatrix} \tag{2.19}$$

但し、ここで $g_{i,j}^{(p)}$、$z_i^{(p)}$ は

$$g_{i,j}^{(p)} = \sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)\frac{\xi_{k_i}\xi_{k_j}}{\sigma_{k_p}^2}$$

$$z_i^{(p)} = \sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)o_{t_p}\frac{\xi_{k_i}}{\sigma_{k_p}^2}$$

である。全ての行 $(1 \le p \le n)$ に関して、式 (2.19) を用いて要素を算出することで変換行列が得られる。

3 変換行列の種類

変換行列のパラメータ数が認識性能に与える影響を調べるために、いくつかの変換行列を用いた。以下に使用した変換行列を示す。w は推定するパラメータ、空白部は推定しないパラメータを示す。また、各行列で 1 列目 ($[w_{1,1}\cdots w_{n,1}]^T$) は offset 項を表す。

- full matrix

$$\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n+1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,n+1} \end{pmatrix}$$

- 対角行列

$$\begin{pmatrix} w_{1,1} & w_{1,2} & & \\ \vdots & & \ddots & \\ w_{n,1} & & & w_{n,n+1} \end{pmatrix}$$

- ブロック対角行列

  同じ物理量の分析パラメータをブロックとする。例えば、ケプストラム分析パラメータが m 次元、Δケプストラムパラメータが $n-m$ 次元ならば以下のブロック対角行列となる。

$$\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,m+1} & & & \\ \vdots & \vdots & \ddots & \vdots & & & \\ \vdots & w_{m,2} & \cdots & w_{m,m+1} & & & \\ \vdots & & & & w_{m+1,m+2} & \cdots & w_{m+1,n+1} \\ \vdots & & & & \vdots & \ddots & \vdots \\ w_{n,1} & & & & w_{n,m+2} & \cdots & w_{n,n+1} \end{pmatrix}$$

- 対角近傍要素行列 (近傍数 k)

  対角要素の近傍に制限した変換行列。 $k=m-1$ の場合ブロック対角行列となる。

$$\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,2+k} & & & & & & & \\ \vdots & \vdots & \ddots & & \ddots & & & & & & \\ \vdots & w_{1+k,2} & & \ddots & & w_{m-k,m+1} & & & & & \\ \vdots & & \ddots & & \ddots & \vdots & & & & & \\ \vdots & & & w_{m,m+1-k} & \cdots & w_{m,m+1} & & & & & \\ \vdots & & & & & & w_{m+1,m+2} & \cdots & w_{m+1,m+2+k} & & \\ \vdots & & & & & & \vdots & \ddots & & \ddots & \\ \vdots & & & & & & w_{m+1+k,m+2} & & \ddots & & w_{n-k,n+1} \\ \vdots & & & & & & & \ddots & & \ddots & \vdots \\ w_{n,1} & & & & & & & & w_{n,n+1-k} & \cdots & w_{n,n+1} \end{pmatrix}$$

4 データ量によるパラメータ数制御方式

線形重回帰モデルを用いた話者適応では、回帰行列のパラメータ数と共有化するクラス数によって認識性能が変化する。また、最適な認識性能を与える回帰行列のパラメータ数と共有化するクラス数は適応データ量に依存する。一般に、適応データ量が少いときはパラメータ数の少い回帰行列を、適応データ量が多いときはパラメータ数の多い回帰行列を使用したときに最適な認識性能が得られると考えられる。そこで、データ量に応じて最適な回帰行列のパラメータ数と共有化するクラス数を決定する方法として、次の 2 つの方法を考えた。

- 提案法 1

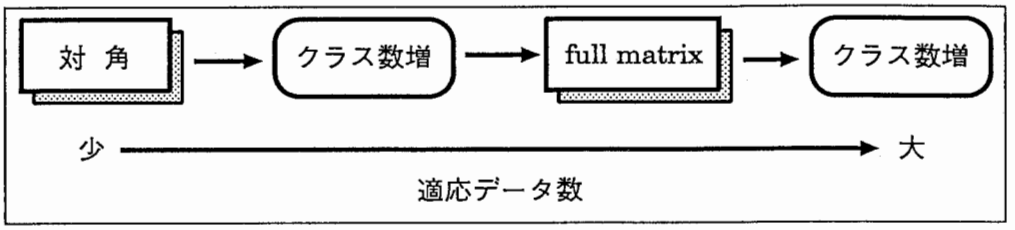

図 2.3: データ量によるパラメータ数制御方式 (提案法 1)

- 提案法 2

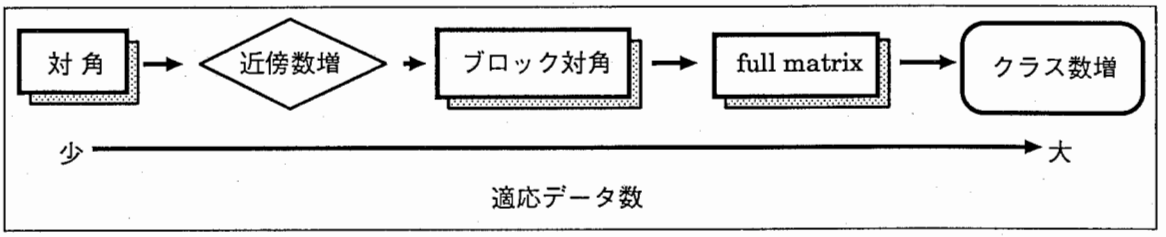

図 2.4: データ量によるパラメータ数制御方式 (提案法 2)

行列とクラス数の切替えには適応データ数による閾値を設定し、それにより切替を行なう。

第 3 章

評価実験

1　実験条件

評価実験として 26 音素を対象とした言語制約を用いない音素タイプライター型の音素認識実験を行なった。表 3.1に音響分析条件を、表 3.2に、使用した音声データを示す。適応前の音響モデルとして不特定話者 (SI) モデルを使用した。この Topology は逐次状態分割法 [4] によって生成した 200 状態各 5 混合の HMnet に 1 状態 10 混合の無音モデルを付加したものを用いた。SI モデルの作成は 285 人分のモデルを混合、合成することによって行なった [5]。実験では各適応文節数に対して選択文節を変えた評価をそれぞれ 3 回繰り返し、平均の音素認識率を求めた。共有化クラスは Kullback Divergence の距離を用い、LBG アルゴリズムによるクラスタリングによって決定した。クラス数を 1、2、4、8(対角は 16、32 まで) とした場合の話者 4 名 (男:2 人 [MAU、MMY]、女:2 人 [FAF、FMS]) の平均音素誤り率で評価した。

実験は大きく分けて 4 つ行なった。以下に実験項目を示す。

1. 変換行列の特性の評価
    - offset の効果の評価
    - ケプストラムとパワーの相関関係の評価
2. データ量とパラメータ数の認識性能への影響評価
    - 適応データ量の変化に対する認識性能
    - 変換パラメータ数の変化に対する認識性能
3. データ量によるパラメータ数制御方式の検討
    - データ量によるパラメータ数制御方式の評価
4. 最大事後確率推定法 (MAP)[6]、最尤推定法 (ML)、移動ベクトル場平滑化法 (VFS)[7] との比較

表 3.1: 分析条件

サンプリング周波数	12KHz
フレーム長	20ms (Hamming window)
フレーム周期	5ms
特徴ベクトル	16 次 LPC ケプストラム
	16 次 Δ ケプストラム
	log パワー
	Δlog パワー

表 3.2: 音声データ

HMnet の生成	男性 1 名 (MHT) の 2640 単語
SI モデル	男性 146 名、女性 139 名 (各話者 50 文章)
評価	279 文節 (SB3 タスク)
適応	598 文節 (SB1,SB2,SB4 タスク) から選択
	1, 5, 10, 20, 30,
	40, 50, 100, 150, 200 文節

2 実験結果

変換行列の特性を調べるために、offset の効果とケプストラムとパワーの相関関係を評価した。

offset の効果の評価結果

offset を用いた時、用いないときの音素誤り率を表 3.3に示す。評価には full matrix を用い、結果には各データ数において一番良い認識率を示したクラスの値を載せた (() 内はクラス数)。offset を用いた時、用いないときを比較して、両方ともほとんど性能に差がないことが分かる。適応データ数が 1 と 5 のところでは、offset なしの方が誤り率が少ないが、これは offset なしの方が行列のパラメータ数が少ないためである。offset の効果がないことになるが、これは使用しているデータがきれいなためで、加算性雑音や回線特性の変動がある場合には offset の効果が現れてくると考えられる。

これ以後の実験では全ての行列で offset を用いたものを使用している。

表 3.3: offset あり、なしの音素誤り率 (%)

適応データ数	offset あり	offset なし
1	69.38(8)	67.16(8)
5	29.95(1)	25.66(1)
10	18.10(1)	18.05(1)
20	16.02(2)	16.08(2)
30	14.74(4)	14.72(4)
40	13.89(4)	13.98(4)
50	13.61(8)	13.60(8)
100	12.06(8)	12.03(8)
150	11.73(8)	11.71(8)
200	11.35(8)	11.34(8)

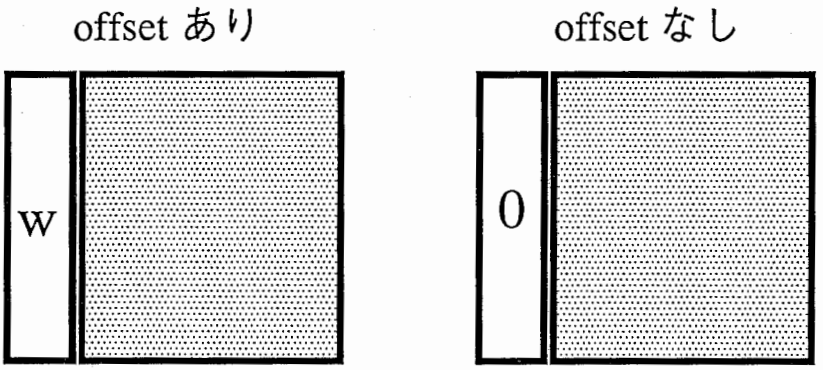

図 3.1: offset あり、 offset なしの変換行列

ケプストラムとパワーの相関関係の評価結果

パワーを独立した時 (4 ブロック対角)、独立しない時 (2 ブロック対角) の音素誤り率を表 3.4に示す。結果には、各データ数において一番良い認識率を示したクラスの値を載せた。パワーを独立した時、独立しない時を比較して、適応データ数が少ない時は独立した方が、適応データ数が多い時は独立しない方が良い結果が得られている。これは、行列のパラメータ数によるものであり、この実験ではケプストラムとパワーの相関関係は確認されなかった。

ブロック対角行列はパラメータ数を減らすことを目的として用いられているので、これ以後の実験ではブロック対角行列にはパワーを独立したものを使用している。

表 3.4: パワーを独立する、しない時の音素誤り率 (%)

適応データ数	パワー独立	独立しない
1	62.88(1)	65.05(8)
5	21.00(1)	21.18(1)
10	18.17(2)	17.73(1)
20	16.01(4)	16.23(4)
30	15.21(8)	15.17(4)
40	14.53(8)	14.63(8)
50	13.96(8)	14.07(8)
100	13.31(8)	13.23(8)
150	13.41(8)	13.07(8)
200	13.10(8)	12.65(8)

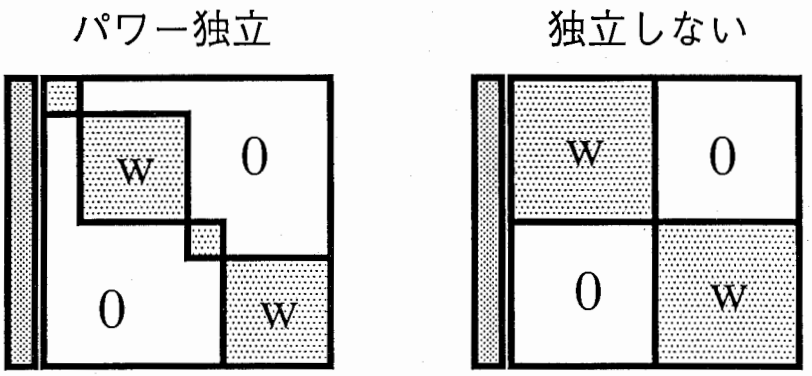

図 3.2: power 独立、非独立の変換行列

2.1 適応データ数による評価結果

図 3.3から 3.7に各変換行列において適応データ数を変化させたときの音素誤り率を示す。図 3.3から full matrix では少量適応データ時の認識性能の低下が大きいことが分かる。また、図 3.4から対角行列では適応データ量が極少いときでも適応の効果が表れているが、適応データ量が多いときでも認識性能の向上が大きくないことが分かる。
図 3.3から 3.7より、

- 適応データ量 少 ⇒ 変換行列数 少
- 適応データ量 大 ⇒ 変換行列数 大

のときに一番良い認識率が得られていることが分かる。

図 3.8で各変換行列において各データ数で一番良い認識率を示したクラスの値によって比較した。図 3.8より、

- 適応データ量 少 ⇒ 変換パラメータ数の少い行列
- 適応データ量 大 ⇒ 変換パラメータ数の多い行列

のときに一番良い認識率が得られていることが分かる。

これらの結果から、最適な認識性能を得るためには、データ量の増加に応じて行列数を多く、パラメータ数の多い行列へ切替えれば良いことが確認できた。

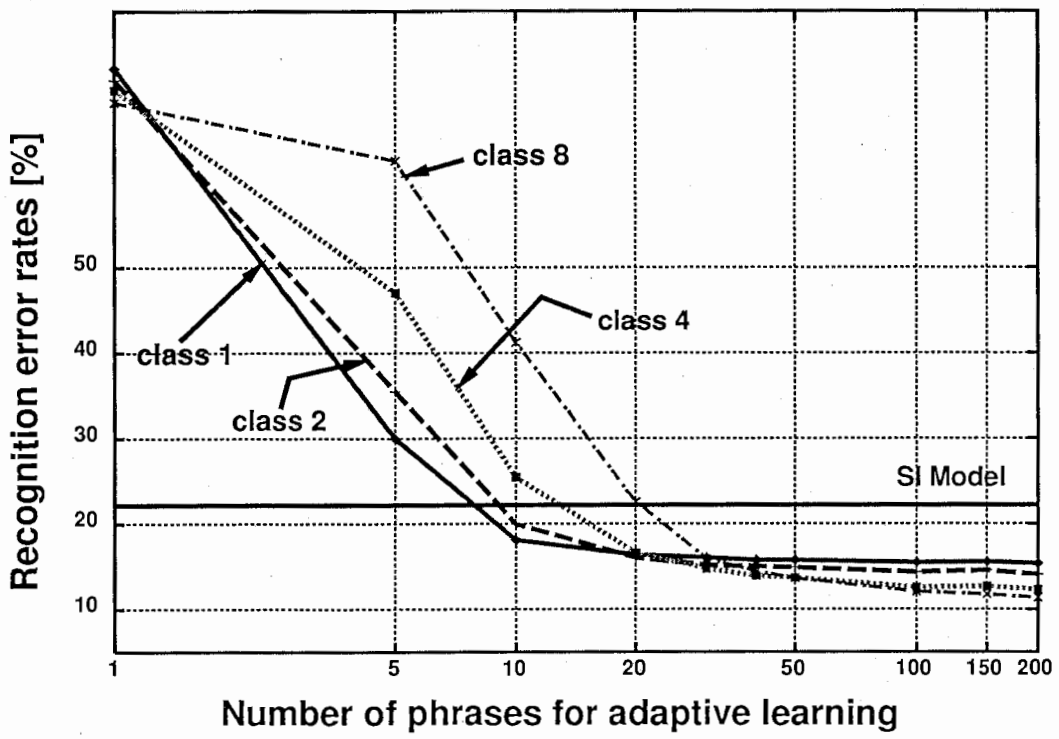

図 3.3: full matrix を用いた時の音素誤り率

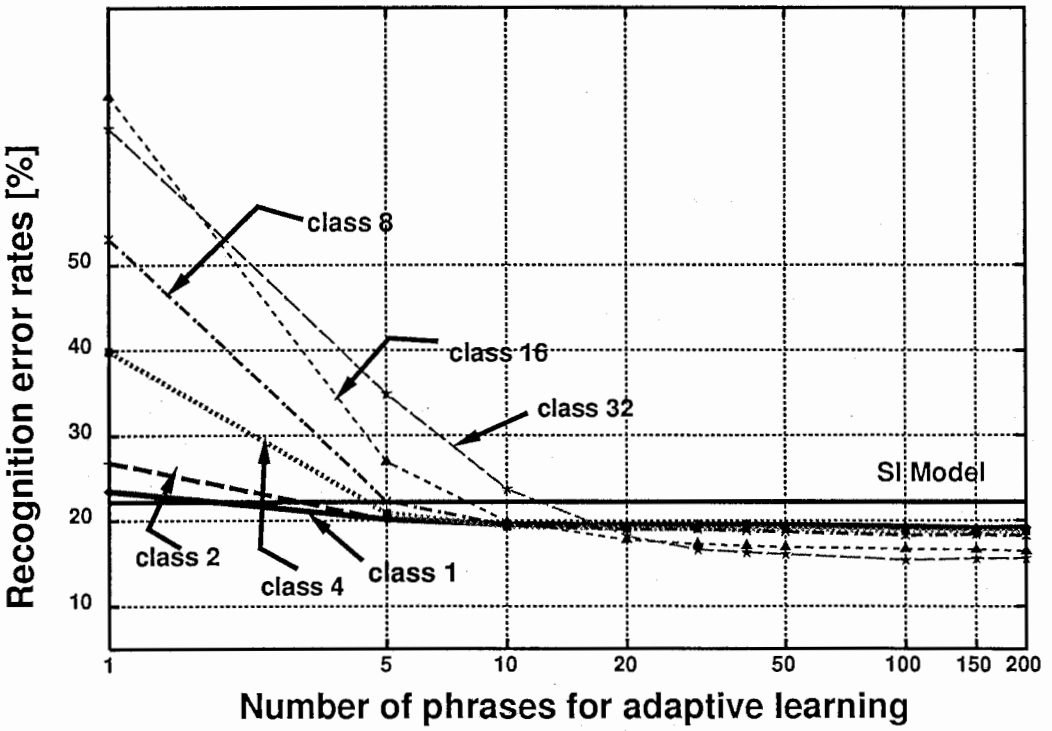

図 3.4: 対角行列を用いた時の音素誤り率

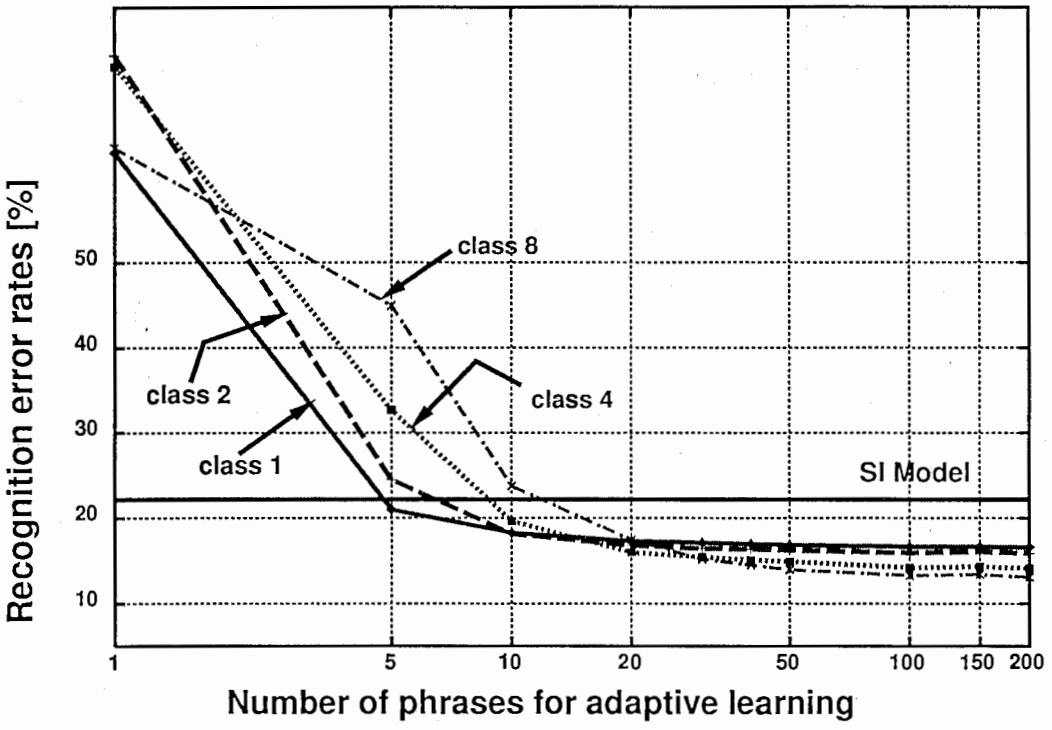

図 3.5: ブロック対角行列を用いた時の音素誤り率

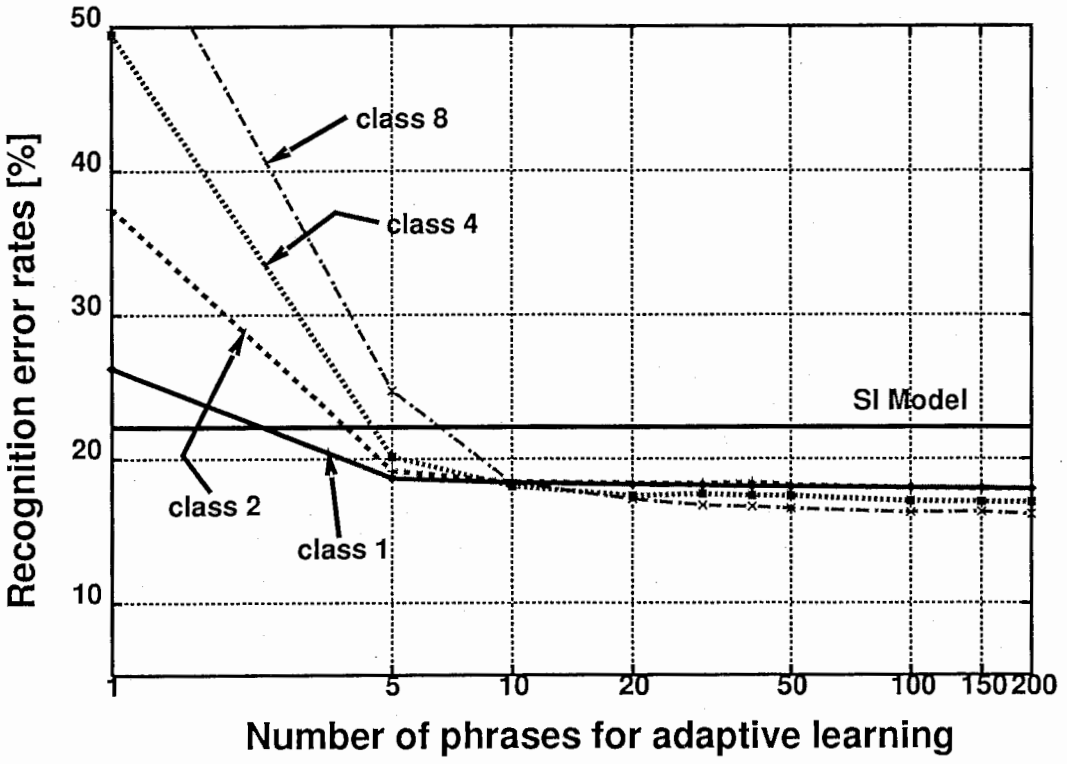

図 3.6: 対角行列 (近傍数 1) を用いた時の音素誤り率

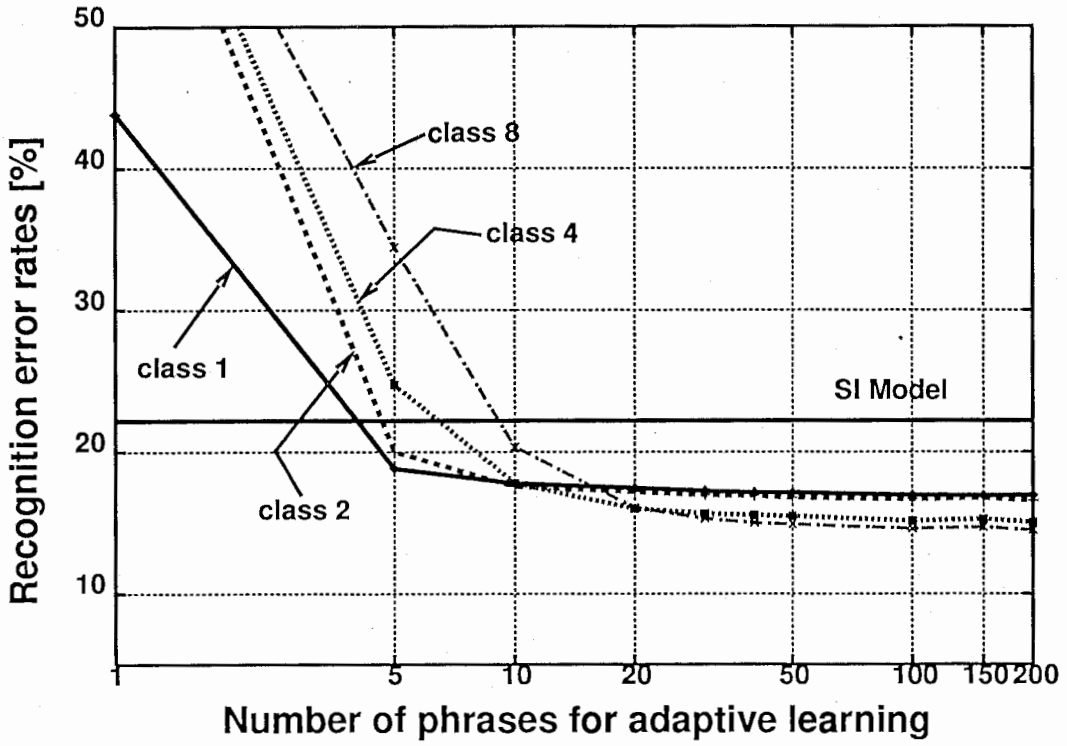

図 3.7: 対角行列 (近傍数 4) を用いた時の音素誤り率

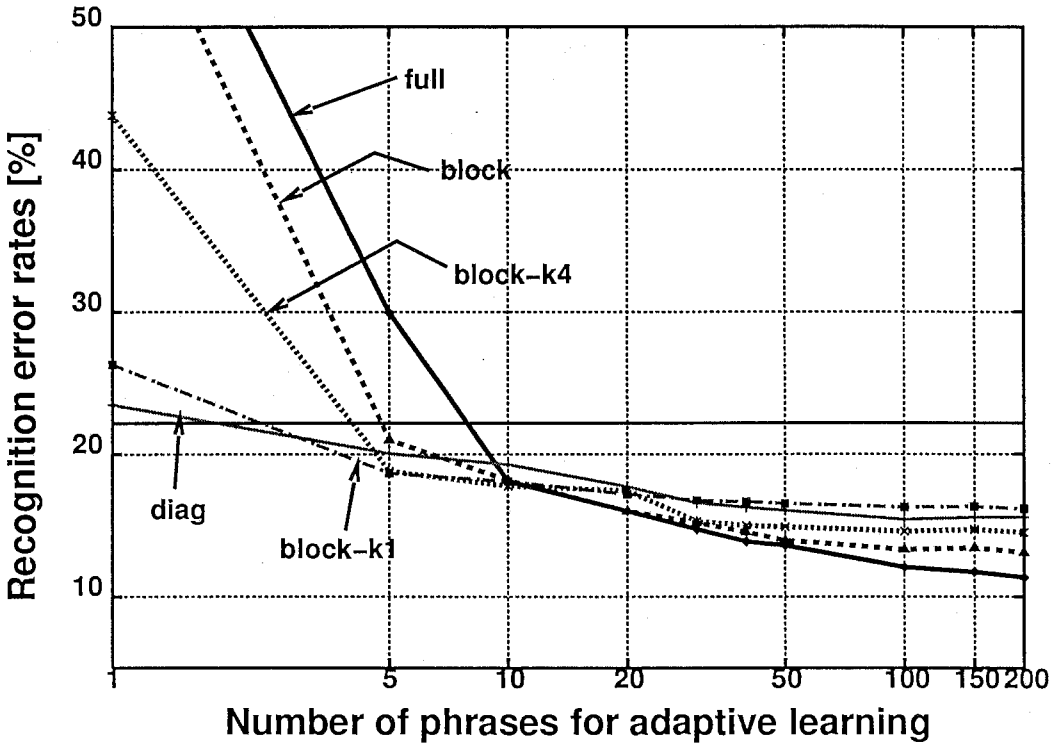
50
40
30
20
10
Recognition error rates [%]
full
block
block-k4
diag
block-k1
1
5
10
20
50
100
150
200
Number of phrases for adaptive learning

図 3.8: 変換行列の比較

2.2 変換パラメータ数による評価結果

行列数とパラメータ数の多い行列への切替をどのように行なえば良いかを考えるために、適応データ数を固定して最適な変換パラメータ数の変化を調べた。文節数が 1、5、10、20、30、40 の時の音素認識実験結果を図 3.9から図 3.14に示す。図 3.9から図 3.14より、
full matrix, ブロック対角では、

- 最適なパラメータ数は適応データの入力フレーム数と同程度

対角行列では、

- 適応データ数少 ⇒ 最適なパラメータ数は適応データの入力フレーム数の 4 〜 5 倍程度
- 適応データ数大 ⇒ 最適なパラメータ数は適応データの入力フレーム数と 2 〜 3 倍程度

であることが分かる。対角近傍要素行列では近傍数が少いときは対角行列と、近傍数が多いときではブロック対角行列と似た傾向が見られた。

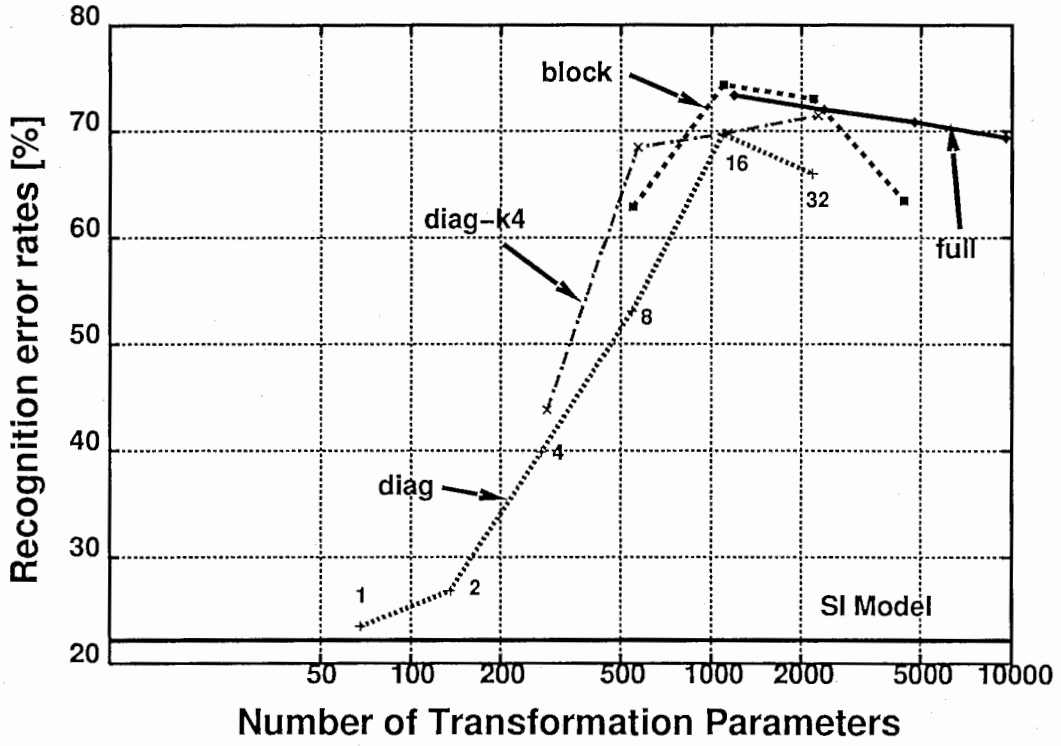

図 3.9: パラメータ数を変化させた時の音素誤り率 (文節数:1(フレーム数:121))

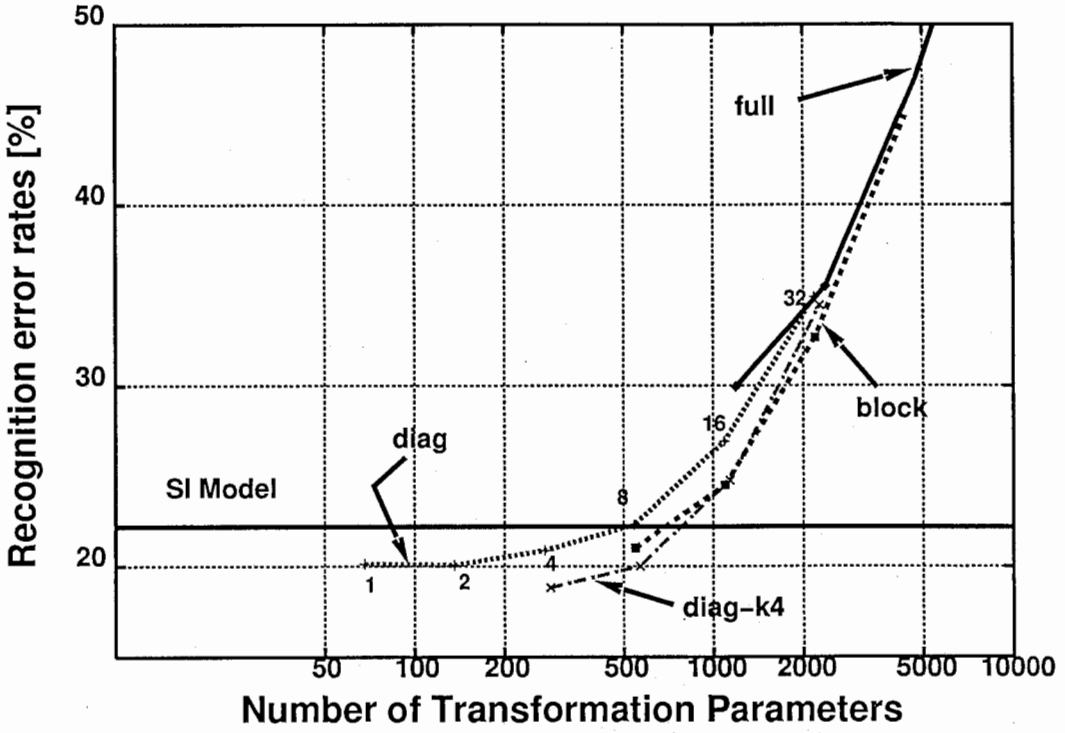

図 3.10: パラメータ数を変化させた時の音素誤り率 (文節数:5⟨ フレーム数:650⟩)

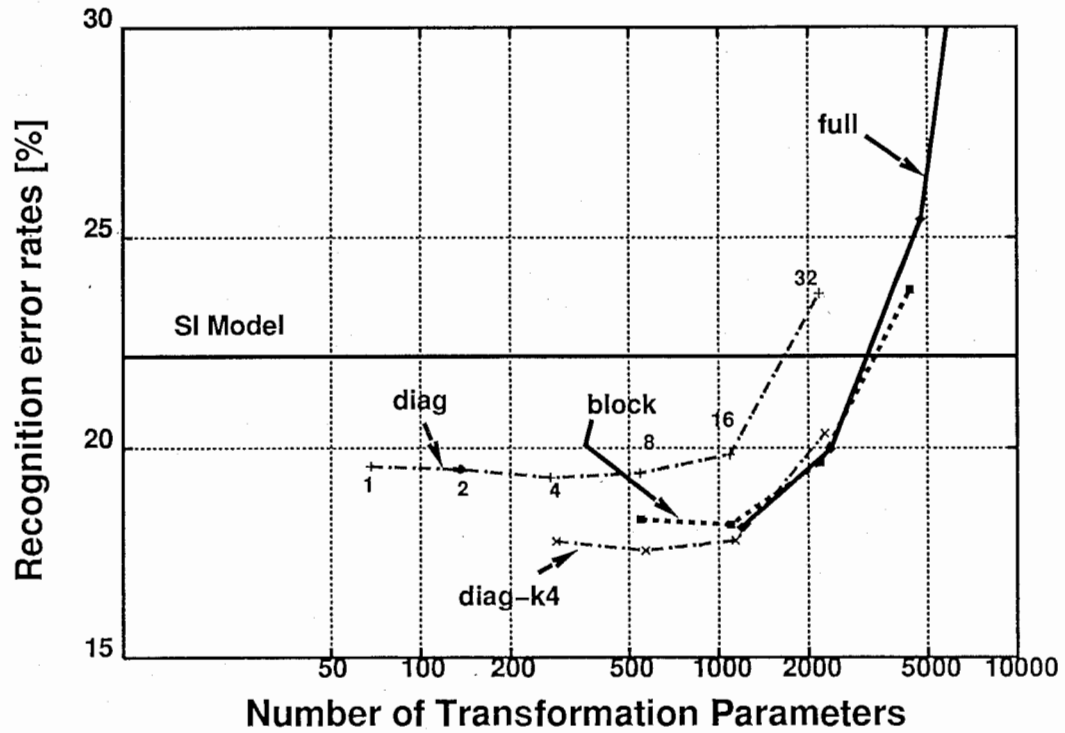

図 3.11: パラメータ数を変化させた時の音素誤り率 (文節数:10⟨ フレーム数:1399⟩)

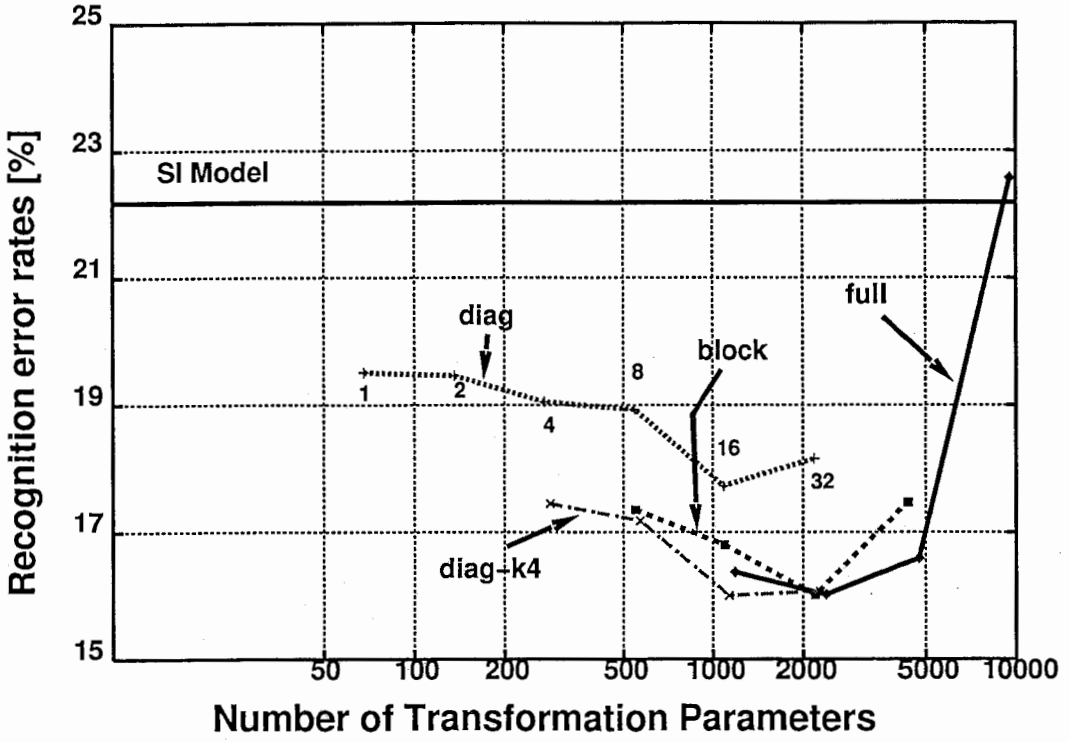

図 3.12: パラメータ数を変化させた時の音素誤り率 (文節数:20⟨ フレーム数:2955⟩)

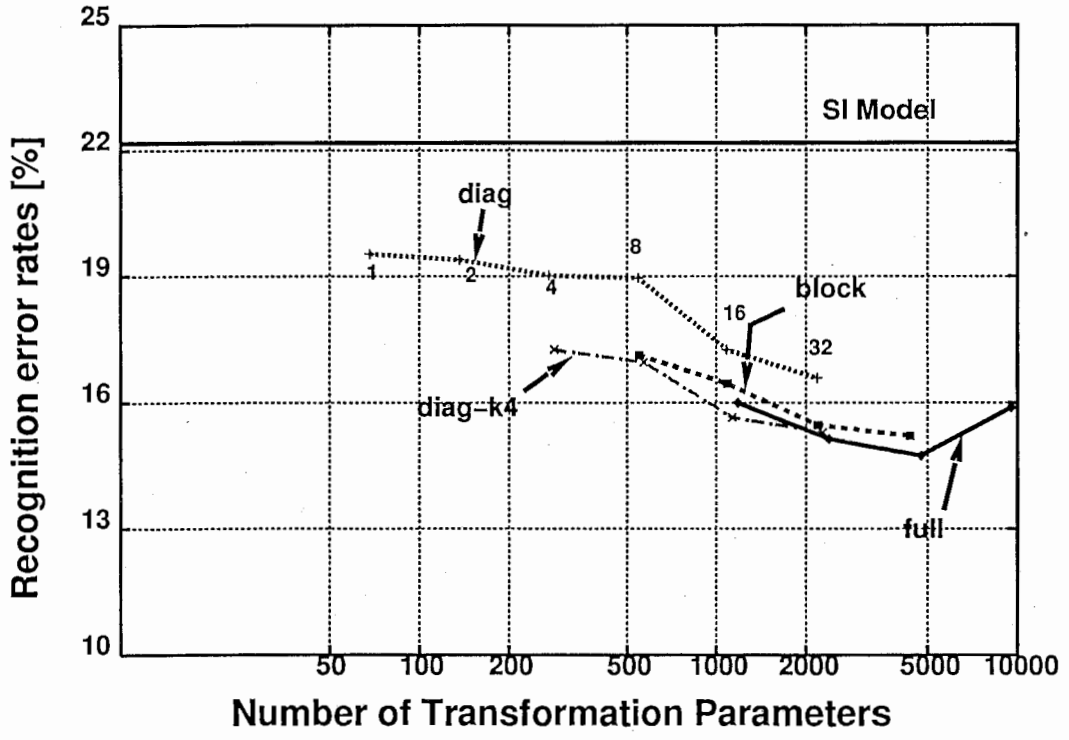

図 3.13: パラメータ数を変化させた時の音素誤り率 (文節数:30⟨ フレーム数:4410⟩)

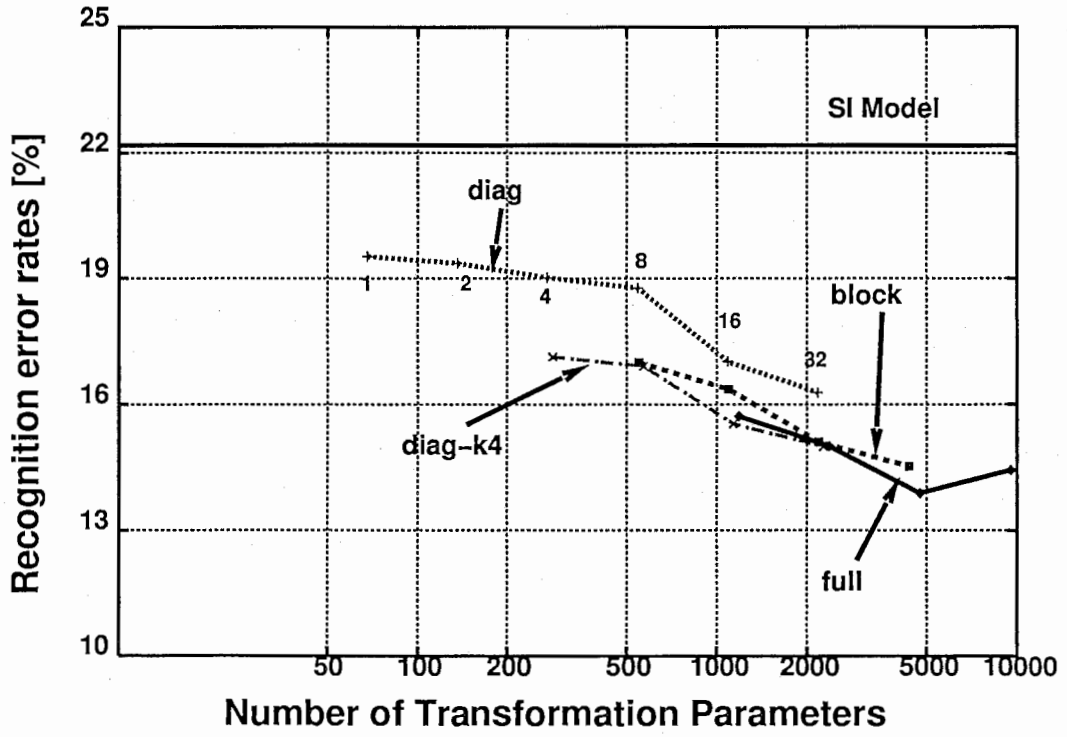

図 3.14: パラメータ数を変化させた時の音素誤り率 (文節数:40〈 フレーム数:5922〉)

2.3 データ量によるパラメータ数制御方式の評価結果

図 3.15に提案する方法と full matrix の音素誤り率を示す。提案法で切替えには実験的に得られた閾値を設定した。図 3.15から、提案する方法では少量の適応データで適応の効果が表れ、full matrix を使用した時の少量の適応データでの認識性能の低下を改善していることが確認できる。提案法 1 と提案法 2 を比較すると提案法 2 が文節数 5 のところで 1% 以上認識率が良く、提案法 2 の対角近傍が有効であることが確認できる。

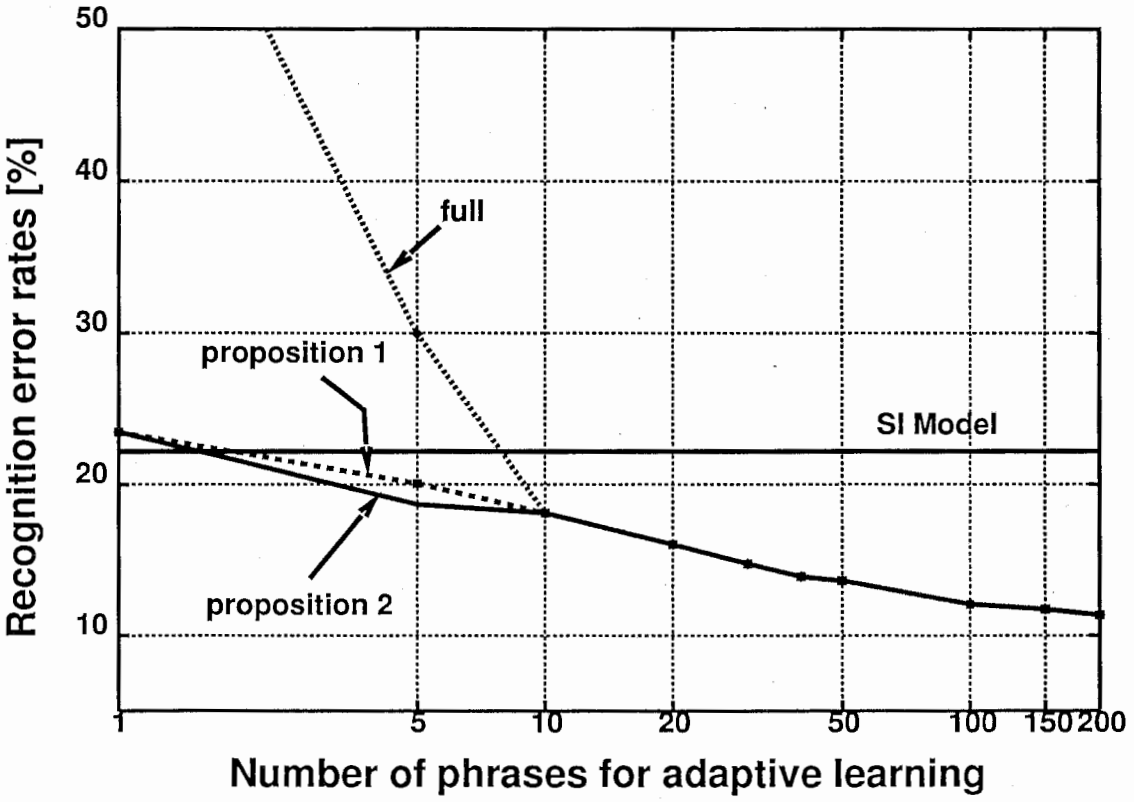

図 3.15: データ量によるパラメータ数制御の音素誤り率

2.4 他の話者適応方式との比較

図 3.16に最尤 (ML) 推定、 MAP 推定、 VFS とデータ量によるパラメータ数制御方式の提案法 2 の音素誤り率を示す。提案法 2 は他の方式と比較して、適応データ量が少ない (1 ～ 50 文節) 領域で高い認識率を示している。提案法 2 では結びを行なうことにより変換行列の共有化を行なっているので、推定するパラメータ数が MAP 推定に比べて少ない。従って、適応データが少量の場合、推定パラメータに対するデータの量が相対的に多くなり、安定した適応が行なわれることにより、高い認識率を示していると考えられる。適応文節数が多い場合は ML、MAP 推定の認識率が高くなるが、これは本研究では全体のデータ量によりパラメータ数制御を行なったためで、各共有化されたクラスに割り当てられるデータ量によりパラメータ数制御を行なうことにより、同等の性能を得ることは可能と考えられる。

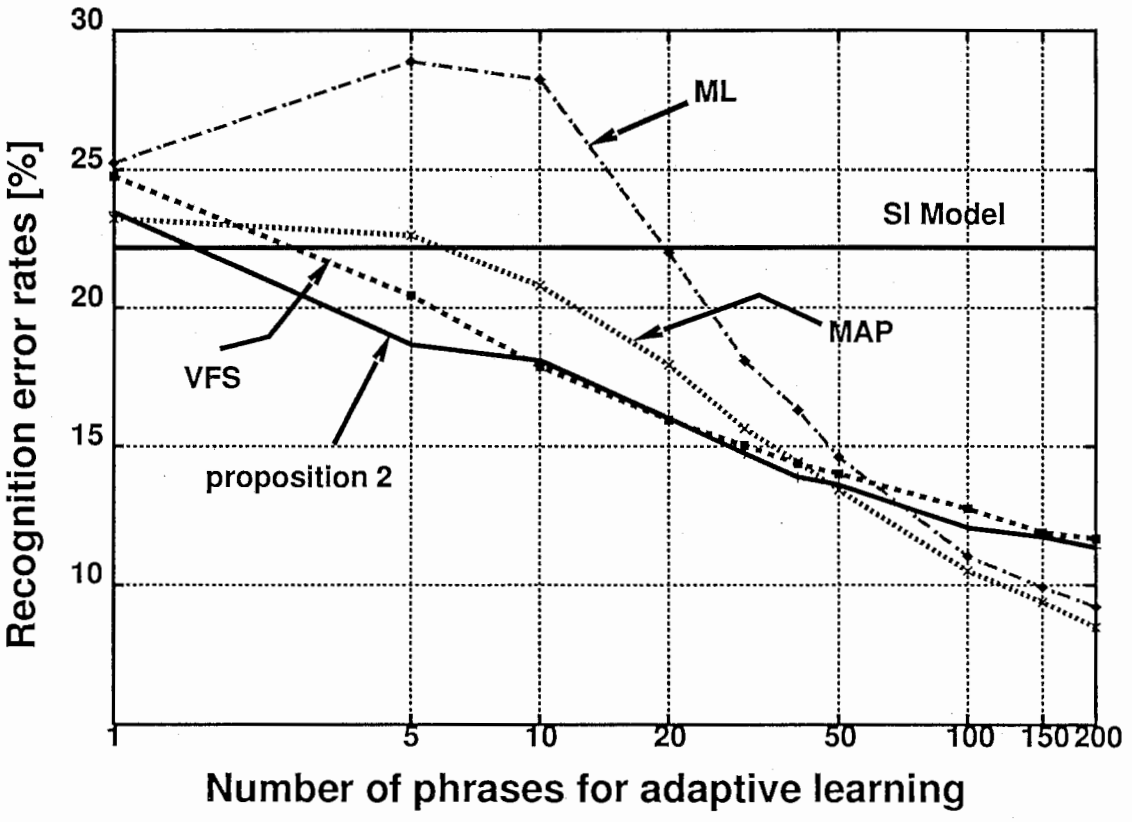

図 3.16: 話者適応方式の比較

第 4 章

結論

本稿では、線形重回帰モデルに基づく話者適応において、認識性能が適応データ量に依存する問題を扱った。MLLR において、適応データ量に応じて回帰行列のパラメータ数と共有化のクラス数を制御する方法により改善を行ない、その有効性を確認した。また、適応データ量により制御した MLLR が少データ量の場合に ML、MAP、VFS よりも高い認識率が得られることを確認した。今後の課題としては、各共有化クラスにおけるデータ量による制御や変換パラメータ数の自動決定などが考えられる。また、本研究では閾値の設定に使用した話者と評価話者が同じであったので、オープンデータでの評価を行なう必要がある。

謝辞

本実習にあたり、親切丁寧な御指導と激励を頂きました、石井 純研究員、匂坂 芳典第一研究室室長、並びにATR音声翻訳通信研究所の皆様に感謝致します。また、本実習の機会を与えて下さった、ATR音声翻訳通信研究所 山崎 泰弘社長、電気通信大学 樽松 明教授に感謝致します。

参考文献

[1] C. L. Leggetter, P. C. Woodland: "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, Vol. 9, pp. 171-185 (1995)

[2] G. Zavaliagkos, R. Schwartz, J. McDonough: "Maximum a Posteriori Adaptation for Large Scale HMM Recognizers," *Proc. of ICASSP 96*, pp. 725-728 (1996)

[3] 茂木、柴田、堀内、松本："制限付重回帰モデルによる話者適応の検討," 音学講論, **1-5-22** (1996.3)

[4] 鷹見、嵯峨山: "音素コンテキストと時間に関する逐次状態分割による隠れマルコフ網の自動生成," 信学技報, **SP91-88** (1991.12)

[5] 小坂、松永、倉岡: "クラスタリング手法を用いた不特定話者モデル作成法," 音学講論,**1-R-12** (1994.11)

[6] C.-H. Lee, C.-H. Lin and B.-H. Juang: "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. on Signal Processing*, vol. 39, no. 4, pp. 806-814 (1991)

[7] 大倉、杉山、嵯峨山: "混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式," 信学技報, **SP92-16** (1992.06)

付録 A

音素認識実験実験結果

1　変換パラメータ数による評価結果

適応文節数 50、100、150、200 のときのパラメータ数の変化による音素誤り率図 A.1から A.4を示す。

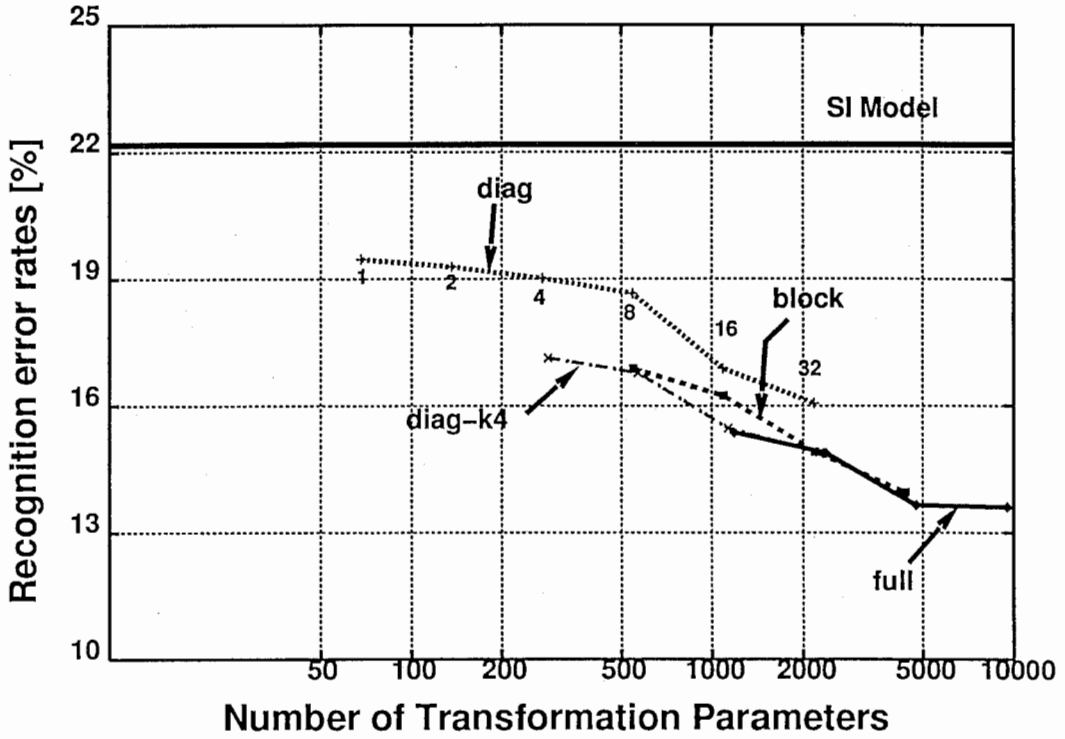

図 A.1: パラメータ数を変化させた時の音素誤り率 (文節数:50〈 フレーム数:7336〉)

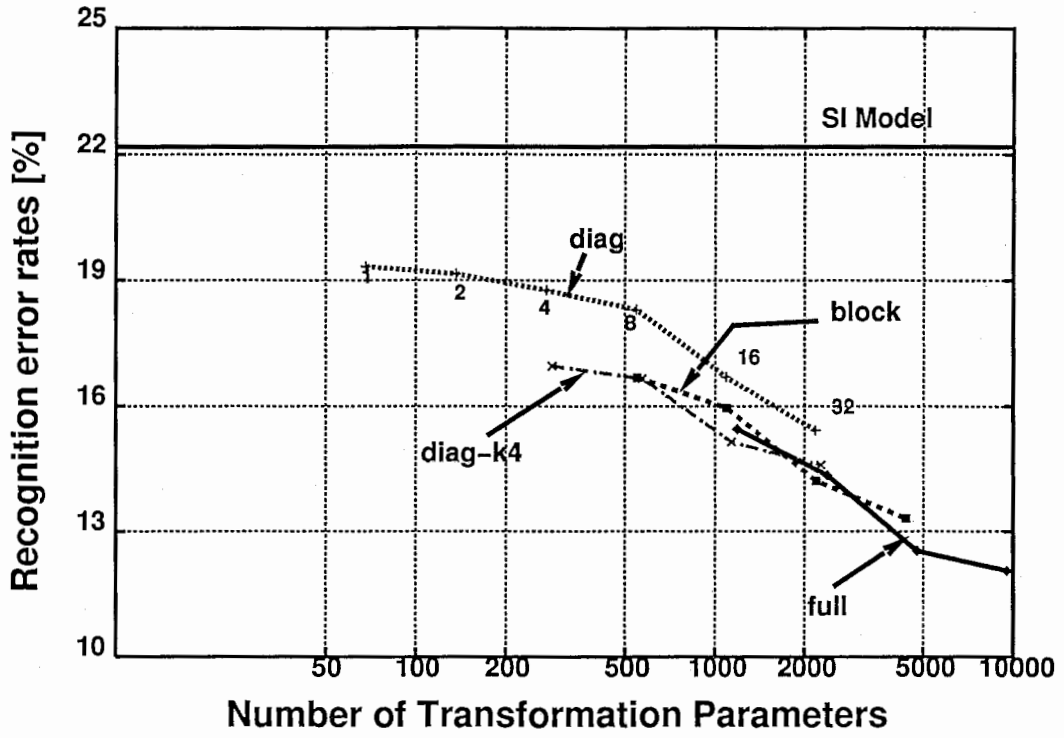

図 A.2: パラメータ数を変化させた時の音素誤り率 (文節数:100⟨ フレーム数:14931⟩)

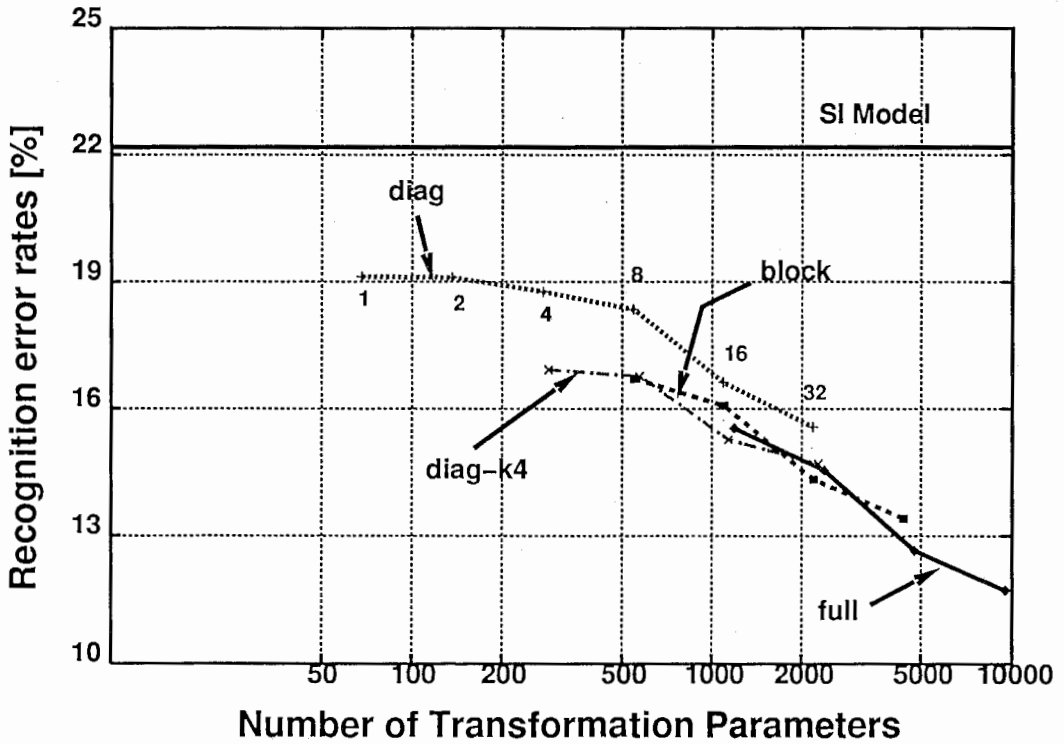

図 A.3: パラメータ数を変化させた時の音素誤り率 (文節数:150⟨ フレーム数:22688⟩)

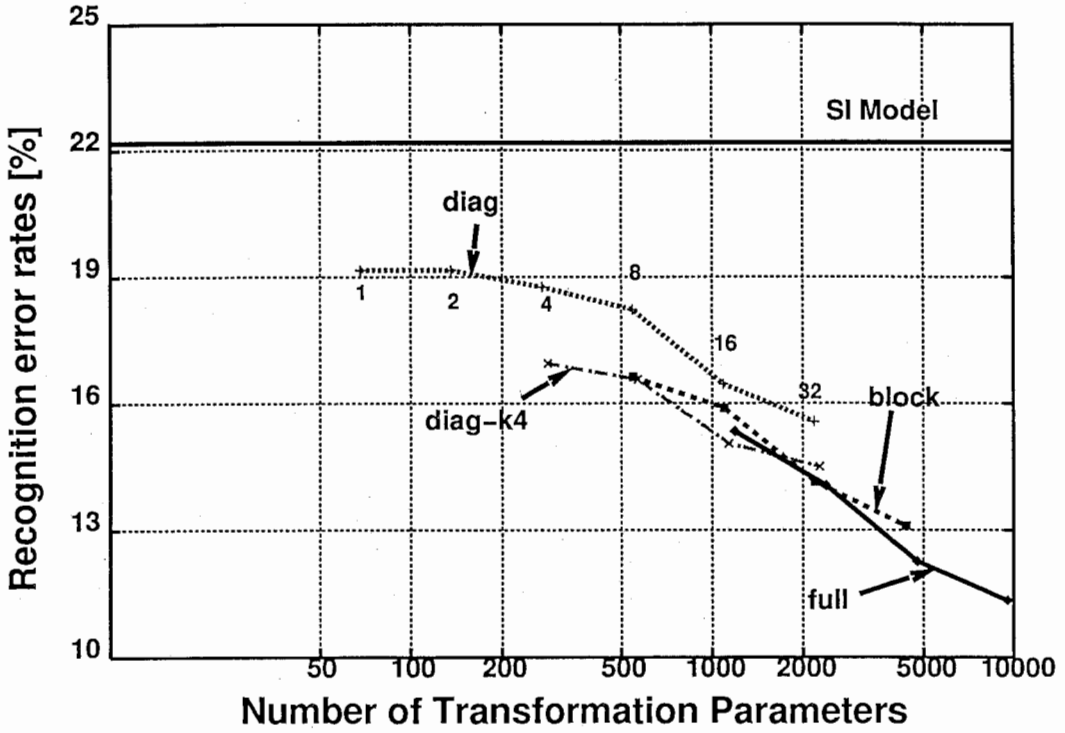

図 A.4: パラメータ数を変化させた時の音素誤り率 (文節数:200〈 フレーム数:30114〉)

2 音素認識実験結果 (各話者)

表 A.1から A.28に各話者毎の全ての音素認識実験結果を示す。

- **MAU**

表 A.1: full matrix の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	73.98	29.74	18.53	15.81	14.88	14.41	14.19	14.26	14.16	14.01
2	72.45	38.73	21.08	15.76	14.09	13.64	13.39	13.02	13.26	12.88
4	68.13	49.32	26.07	16.36	14.07	12.94	12.53	11.48	11.38	11.15
8	65.99	60.82	40.47	21.30	15.20	14.32	13.43	10.64	10.33	10.18

表 A.2: 対角行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	24.35	19.68	18.88	18.41	18.19	18.17	18.04	17.91	17.35	17.34
2	27.75	19.76	19.05	18.79	18.63	18.51	18.28	17.96	17.80	17.89
4	40.98	20.97	19.60	18.89	18.44	18.48	18.38	18.40	18.14	18.26
8	54.79	21.33	19.26	18.59	18.49	18.15	17.65	17.67	17.52	17.30
16	70.00	26.42	20.35	17.31	16.76	16.30	15.90	16.16	15.65	15.58
32	69.56	35.00	24.38	16.84	15.13	14.86	14.44	14.44	14.18	14.18

表 A.3: ブロック対角行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	63.07	23.81	18.75	17.04	16.07	15.95	15.72	15.50	15.37	15.36
2	74.38	26.96	18.79	16.36	15.37	14.81	14.75	14.63	14.65	14.62
4	71.05	34.18	21.12	16.16	15.23	14.27	13.62	12.99	12.87	12.54
8	63.94	47.85	24.97	18.05	15.42	14.26	13.52	12.65	12.39	12.12

表 A.4: 対角 (近傍数 1) 行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	27.01	18.36	17.39	16.95	16.81	16.76	16.51	16.36	16.17	16.02
2	37.15	19.24	17.87	17.74	17.35	17.19	16.97	16.96	16.55	16.70
4	49.62	20.70	18.28	17.14	16.85	16.76	16.66	16.07	16.05	16.15
8	59.98	23.57	18.74	17.09	16.35	16.16	15.97	15.90	15.48	15.51

表 A.5: 対角 (近傍数 4) 行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	43.18	18.61	17.82	16.85	16.40	16.13	15.70	15.41	15.17	15.32
2	65.33	19.69	16.80	15.90	15.51	15.28	15.15	14.58	14.66	14.48
4	68.43	25.67	18.01	14.81	14.65	14.11	13.94	13.76	13.54	13.24
8	59.98	23.57	18.74	17.09	16.35	16.16	15.97	15.90	15.48	15.51

表 A.6: full matrix(offset なし) の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	73.53	28.91	18.83	15.86	15.02	14.38	14.14	14.17	13.84	13.64
2	71.99	38.88	21.09	15.83	14.38	13.51	13.64	12.96	13.18	13.02
4	68.29	49.21	25.64	16.40	14.36	13.19	12.64	11.14	11.32	11.04
8	72.19	63.89	40.62	21.39	15.47	14.41	13.43	10.66	10.35	10.21

表 A.7: ブロック対角行列 (パワー独立しない) の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	64.96	24.06	18.49	16.49	15.73	15.47	15.23	15.48	15.15	14.98
2	74.44	28.20	19.18	16.72	15.10	14.39	14.42	14.21	14.13	14.12
4	72.27	35.65	22.02	16.67	14.62	13.98	13.64	12.99	12.87	12.67
8	64.20	49.17	25.26	17.67	14.97	13.98	13.42	12.40	11.84	11.68

- MMY

表 A.8: full matrix の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	70.12	25.79	19.78	17.63	17.48	17.20	17.24	17.18	17.53	17.21
2	70.64	32.10	20.75	16.81	16.36	16.50	16.52	16.17	16.33	15.81
4	74.46	44.78	25.27	16.72	15.63	15.37	14.90	13.92	13.97	13.45
8	73.29	62.01	39.88	22.63	16.97	15.39	14.33	13.08	12.76	12.27

表 A.9: 対角行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	20.27	18.70	18.82	18.87	19.07	19.23	19.20	18.94	18.85	19.07
2	23.26	19.19	19.10	19.24	19.59	19.54	19.59	19.06	19.15	19.16
4	37.27	20.62	19.11	18.65	19.11	19.25	19.33	18.77	18.79	18.71
8	49.38	21.93	19.33	18.89	19.39	19.15	19.01	18.27	18.40	18.31
16	70.50	25.98	20.51	18.08	18.07	17.83	17.57	17.39	17.15	17.04
32	64.40	31.64	24.39	19.46	17.78	17.43	17.20	16.58	16.46	16.62

表 A.10: ブロック対角行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	60.30	20.32	19.01	18.18	18.27	18.27	18.12	17.87	17.82	17.54
2	69.06	23.65	19.18	17.68	17.53	17.67	17.52	17.05	16.91	16.72
4	72.01	31.03	19.79	17.08	16.86	16.56	16.53	15.50	15.65	15.84
8	63.11	44.36	24.23	18.32	16.74	16.01	15.44	14.74	14.91	14.47

表 A.11: 対角 (近傍数 1) 行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	23.02	17.94	18.52	18.41	18.92	18.57	18.57	18.60	18.22	18.31
2	30.88	18.53	18.66	18.45	18.81	18.81	18.68	18.50	18.45	18.47
4	44.57	19.75	18.51	17.87	18.40	18.23	18.43	18.42	18.37	18.31
8	63.30	23.10	19.53	18.04	18.31	18.21	18.09	17.80	18.08	17.80

表 A.12: 対角 (近傍数 4) 行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	41.05	19.65	18.65	18.18	18.35	18.22	18.45	18.09	17.74	17.77
2	56.03	21.00	18.56	17.73	17.72	17.47	17.49	17.34	17.60	17.20
4	65.90	24.07	19.16	17.04	16.70	16.69	16.66	16.74	17.06	16.77
8	63.30	23.10	19.53	18.04	18.31	18.21	18.09	17.80	18.08	17.80

表 A.13: full matrix(offset なし) の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	70.01	25.45	19.54	17.64	17.33	17.03	17.20	17.16	17.40	17.11
2	70.67	31.21	20.60	16.91	16.69	16.72	16.69	16.56	16.53	15.78
4	67.19	45.00	24.62	16.62	15.38	15.08	14.94	14.18	14.37	13.78
8	65.37	62.42	39.72	22.44	16.95	15.65	14.18	13.10	12.94	12.30

表 A.14: ブロック対角行列 (パワー独立しない) の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	62.09	20.60	18.97	17.98	18.46	18.08	18.17	17.78	17.78	17.64
2	71.48	24.46	19.33	17.72	17.25	17.34	17.10	16.66	16.47	16.40
4	71.75	33.08	20.41	17.18	16.60	16.70	16.55	15.77	15.63	15.33
8	65.54	45.35	24.93	18.70	16.82	16.02	15.48	14.66	14.38	14.02

- **FAF**

表 A.15: full matrix の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	75.54	24.37	15.61	15.08	15.08	15.13	15.21	14.90	15.09	15.00
2	73.62	34.56	17.38	14.17	13.76	13.90	13.71	13.42	13.72	13.45
4	71.58	44.95	23.51	15.16	13.06	12.46	12.24	10.96	11.40	11.07
8	73.77	61.68	41.03	21.59	14.05	12.29	11.98	11.15	10.59	10.36

表 A.16: 対角行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	24.41	20.58	19.91	20.05	19.94	20.04	20.00	19.95	19.74	19.77
2	28.03	20.88	20.29	20.29	19.77	19.94	19.91	19.94	19.81	19.82
4	39.84	20.34	19.14	19.26	19.05	18.86	18.71	18.66	18.85	18.90
8	56.13	21.50	18.65	19.20	19.09	19.09	19.07	18.72	18.85	18.81
16	70.36	25.64	18.32	17.56	16.93	17.01	17.03	16.84	17.11	16.92
32	65.75	35.39	20.99	17.88	16.57	15.89	15.79	15.13	15.66	15.64

表 A.17: ブロック対角行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	69.29	18.96	16.18	15.88	16.00	16.09	16.04	16.20	16.48	16.49
2	78.83	22.73	15.98	15.46	15.33	15.31	15.36	15.48	15.86	15.65
4	74.31	32.44	17.52	14.24	13.98	13.97	13.86	13.28	13.50	13.60
8	63.45	41.10	21.59	15.46	13.27	12.84	12.34	11.63	11.90	11.89

表 A.18: 対角 (近傍数 1) 行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	25.65	18.10	17.59	18.01	17.86	17.98	17.97	17.89	17.96	17.94
2	39.96	18.35	17.21	17.87	17.72	18.07	17.91	17.83	17.78	17.79
4	51.79	18.47	16.43	16.44	16.77	16.85	16.44	15.98	16.03	16.00
8	57.44	24.15	15.56	15.69	15.04	15.14	14.89	14.81	15.26	15.04

表 A.19: 対角 (近傍数 4) 行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	50.45	17.64	16.64	17.07	17.01	17.12	17.21	17.43	17.71	17.62
2	69.59	18.70	16.14	16.90	16.99	17.14	16.95	17.23	17.21	17.26
4	76.67	23.91	15.93	15.31	14.94	15.16	14.94	14.54	14.76	14.81
8	57.44	24.15	15.56	15.69	15.04	15.14	14.89	14.81	15.26	15.04

表 A.20: full matrix(offset なし) の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	74.89	23.30	15.65	15.16	15.03	15.59	15.44	15.05	15.24	15.24
2	73.13	34.35	17.21	14.32	13.96	13.98	14.12	13.72	13.98	13.91
4	68.02	44.91	23.42	15.29	13.17	12.59	12.32	11.25	11.35	11.19
8	66.60	61.49	40.16	21.35	14.32	12.53	12.04	10.96	10.56	10.45

表 A.21: ブロック対角行列 (パワー独立しない) の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	73.03	19.60	15.76	15.81	15.59	15.88	15.70	15.78	15.90	15.88
2	77.97	23.17	15.78	14.94	14.94	15.50	15.28	15.15	15.76	15.65
4	73.71	34.05	18.42	14.64	13.81	13.77	13.58	12.91	13.52	13.55
8	64.72	43.41	23.11	15.96	13.47	12.89	12.49	11.73	11.65	11.53

- **FMS**

表 A.22: full matrix の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	73.80	24.90	18.47	16.94	16.55	16.14	16.29	15.47	15.38	15.15
2	71.32	36.59	20.67	17.32	16.35	15.98	15.90	14.78	14.84	14.11
4	69.26	49.11	26.89	18.13	16.20	14.78	14.96	13.80	13.89	13.41
8	64.46	65.70	43.47	24.76	17.32	15.77	14.68	13.39	13.26	12.59

表 A.23: 対角行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	24.85	21.58	20.65	20.71	20.92	20.65	20.62	20.52	20.52	20.46
2	28.33	20.40	19.52	19.53	19.58	19.44	19.36	19.63	19.66	19.79
4	41.40	21.50	19.32	19.42	19.47	19.47	19.66	19.26	19.32	19.17
8	52.03	24.35	20.37	19.04	18.87	18.71	18.89	18.53	18.64	18.49
16	67.99	29.60	20.20	17.98	17.28	16.92	17.07	16.40	16.62	16.30
32	64.28	37.54	24.92	18.48	16.87	16.94	16.79	15.54	15.99	15.82

表 A.24: ブロック対角行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	58.87	20.91	19.23	18.29	18.19	17.71	17.63	17.13	17.17	17.12
2	75.14	24.62	18.72	17.71	17.59	17.63	17.33	16.67	16.87	16.57
4	74.77	33.02	20.18	16.55	15.80	15.65	15.63	15.08	15.30	14.59
8	63.38	46.47	24.24	18.03	15.42	15.02	14.54	14.23	14.46	13.93

表 A.25: 対角 (近傍数 1) 行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	29.48	20.27	19.96	19.51	19.18	19.03	19.28	19.14	19.52	19.41
2	41.47	20.52	19.92	19.17	19.18	19.37	19.07	18.67	18.87	18.76
4	53.94	21.68	19.18	18.28	18.16	17.97	18.16	17.78	17.61	17.52
8	59.28	28.07	19.49	17.99	17.39	17.19	17.23	16.63	16.44	16.32

表 A.26: 対角 (近傍数 4) 行列の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	40.55	19.38	17.99	17.71	17.27	17.05	17.14	16.93	17.10	17.07
2	61.87	20.62	18.68	18.17	17.62	17.74	17.52	17.51	17.61	17.34
4	68.31	25.36	18.04	16.85	16.30	16.18	16.34	15.52	15.74	15.38
8	71.32	36.05	21.05	17.22	16.00	16.14	16.05	15.78	16.04	15.59

表 A.27: full matrix(offset なし) の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	74.29	24.97	18.17	17.07	16.18	16.20	16.28	15.62	15.54	15.10
2	71.76	37.10	20.71	17.27	16.33	16.10	15.99	14.83	14.89	14.24
4	69.44	48.81	26.00	18.03	15.99	15.07	14.88	13.61	13.66	13.10
8	64.46	66.47	43.31	24.27	17.32	15.54	14.76	13.38	13.01	12.39

表 A.28: ブロック対角行列 (パワー独立しない) の音素誤り率 (%)

	適応データ数 (文節数)									
クラス数	1	5	10	20	30	40	50	100	150	200
1	64.39	20.46	17.71	16.95	16.50	16.39	16.20	15.92	15.85	15.28
2	75.28	25.05	17.71	16.79	16.34	16.27	16.18	15.82	15.74	15.25
4	72.30	35.24	20.65	16.45	15.65	16.03	15.79	14.84	15.14	14.50
8	65.73	46.93	25.83	18.54	15.78	15.65	14.88	14.13	14.41	13.38