TR-IT-0022

木構造話者クラスタリング手法を用いた話者適応アルゴリズムにおける諸検討

Investigation of a Speaker Adaptation Based on Tree-structured Clustering Algorithm

嶺 竜治　小坂 哲夫
Ryuji MINE, Tetsuo KOSAKA

概要

木構造を用いて話者モデルを階層的にクラスタリングし、話者適応する手法を提案する。これを音素認識実験をおこなって評価した結果について報告する。また、本手法における諸検討 (1) 話者のクラスタリングに用いる距離尺度の検討 (Bhattacharyya 距離と Kullback 距離)、(2) 話者選択後の話者重み学習による話者適応の検討についても述べる。

目次

表目次

図目次

1 はじめに

近年音声情報処理技術が急速に発展し、認識の分野での研究の関心事は、大語彙、連続音声、不特定話者へと発展しつつある。これらのうち、本研究では、不特定話者の音声認識を行なう時の話者の個人性の問題について扱う。

音声はもともと個人の発声器官の違いや習慣の違いをふくんでおり、きわめてさまざまな形式をもったものである。これを従来の特定話者向けの音声認識システムで認識を行なうというのは自ずから限界があるものと思われる。そこで認識に際してシステムになんらかの対策が必要となる。これを、話者適応と呼んでいる。

このような適応化技術を行なうことは、話者の個人性を吸収できるという利点だけでなく、マイクの特性、周波数雨特性、音響特性、背景雑音などの変動もカバーすることが期待できる。

本研究で用いられる音声認識アルゴリズムは基本的には、多数の話者のデータを用意し、クラスタリング手法を用いて話者の違いによる個人性の変動をカバーするような複数のテンプレートを用意し、テストデータとこれらすべてのテンプレートとの類似度を計算していく、いわゆるマルチテンプレート手法（図 1）に基づいている。

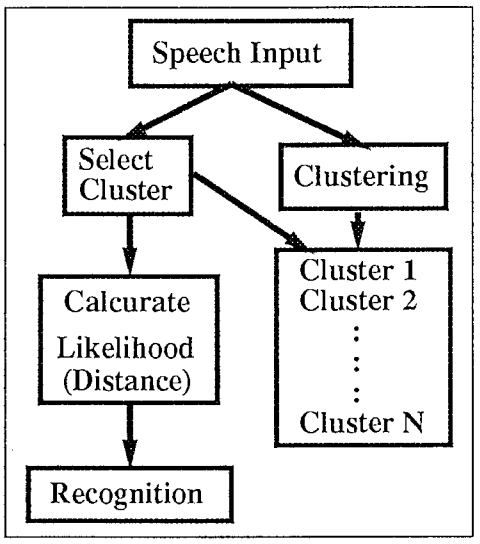

図 1: マルチテンプレート方式による話者適応

しかし、このような手法は大量の話者のテンプレートを用いたシステムには不向きである。またどれだけのテンプレートの話者を用意すべきかは明らかではない。

本研究ではこの問題を解決するため、階層的な話者クラスタリングによる話者適応法を提案する。この方法では話者特性を階層的に逐次分割することにより、話者モデルの木構造を作成する。この木構造を、入力音声に対するモデルの尤度を基準として探索することにより話者選択を行なう。この際に尤度が最大となるノードにおけるモデルを選択することにより、頑健性の低下を防ぐことが可能となる。

また話者モデル選択による話者適応では、性能向上のためには学習用話者数を増加する必要があるが、本方式では木構造で話者クラスタを表現することにより、話者が増加した場合の、適応に要する計算量の増大を防ぐという効果も得られる。

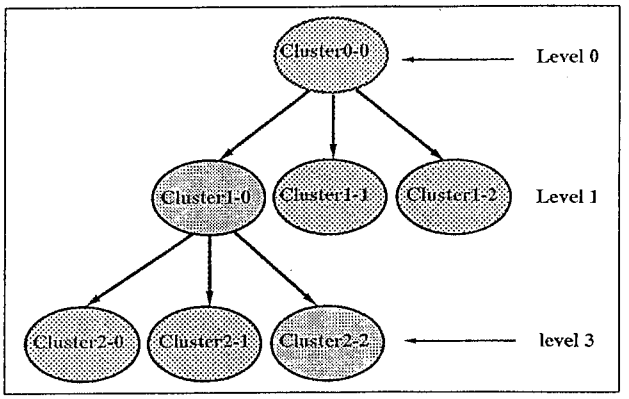

図 2: 話者の木構造

評価実験として、実際に 170 人の話者の音素モデルセットをクラスタリングして木構造を作成し、これに基づき話者適応の実験を行なった。

さらに、この適応アルゴリズムを用いて、(1) クラスタリングの際に用いる話者間の距離尺度に関する検討および (2) 話者結び重みづけ学習法に関する検討を行なった。

2 アルゴリズム

2.1 話者クラスタの木構造の作成

本報告では音素モデルとして Hidden Markov Network (HMnet) を使用した [11]。HMnet は異音間で状態を共有したモデルで、ひとつの HMnet で学習データに現われる全ての音素環境を含む音素環境依存モデルの集合が表現できる。

表 1に木構造作成アルゴリズムを示す。このアルゴリズムでは木構造の各ノードの最大クラスタ数 K のみを与える。木構造の各ノードでのクラスタリングには SPLIT 法で用いられているクラスタリングアルゴリズムを使用した [12]。SPLIT 法によるクラスタリング手法は k-means 法のように初期値を設定する必要はない。また歪みのもっとも大きいところを逐次分割していくため、コードブックサイズは 2 の倍数である必要はない。ただし距離尺度として統計的モデルが扱えるよう Bhattacharyya 距離、CrossEntropy 距離を用いた。

表 1: 木構造作成アルゴリズム

STEP1 $j = 0$。j は木構造の階層を表す。N 人の話者のデータで N 個の HMnet を作成する。全体を 1 クラスタとする。

STEP2 全てのクラスタ l において $S_l(j)$ の話者数が K 未満のときクラスタ l のクラスタリングを終了。但し $S_l(j)$ は l 番目のクラスタに属する話者の集合。

STEP3 STEP2 の終了条件以外の全てのクラスタ l において、$S_l(j)$ を K 個にクラスタリング。各クラスタに属する話者のデータを用いて、HMnet のパラメータ推定を行ない K 個の HMnet $\{m_1^l(j+1), \ldots, m_K^l(j+1)\}$ を作成。

STEP4 $j \leftarrow j+1$。STEP2 へ。

2.2 SPLIT 法によるクラスタリング

SPLIT 法によるクラスタリングアルゴリズムを表 2に示す。このアルゴリズムは、各話者からもっとも近いテンプレートへの距離の平均 (D_{AV}) がしきい値 (D_{TH}) 以下、かつ同数のテンプレートの組合せの中では最小になるようにテンプレートを抽出するという距離最小化の原理に基づいている。

表 2: SPLIT 法によるクラスタリングアルゴリズム

STEP1 N 人の単語音声を収集

STEP2 単語ごとに $N \times N$ の話者間距離 Matrix を作成

STEP3 各話者からの距離が最初となる話者 A を選択。

STEP4 $D_{AV} < D_{TH}$ なら終了

STEP5 2 話者を選択

STEP6 $D_{AV} < D_{TH}$ なら終了

STEP7 各クラスタメンバの更新

STEP8 中心核への距離の和が最も大きいクラスタについて 2 つの中心核を選択

STEP9 各クラスタメンバの更新

STEP10 各クラスタについて各メンバからの距離の平均が最小になる中心核を選ぶ

STEP11 $D_{AV_i} \neq D_{AV_{i-1}}$ なら STEP 9 へ

STEP12 $D_{AV} < D_{TH}$ なら終了、でなければ STEP 8 へ

2.3 話者適応アルゴリズム

話者適応は HMnet の出力尤度に基づいて、クラスタリングによって作られたモデルを選択し、木構造を探索することにより行なう。木構造の探索中に各ノードに属するモデルに対する尤度を記憶しておき、探索の終了後最大尤度を与えるノードのモデルを用いて認識を行なう。アルゴリズムを表 3に示す。STEP3 におけるモデルに対する尤度の計算法として type1・type2 の 2 種類の実験を行なった。type2 は認識と同じ評価基準による選択であるのに対し、type1 は評価基準は異なるが STEP2 の結果をそのまま用いることができるので、計算量が少ないという特徴がある。

表 3: 話者適応アルゴリズム

STEP1 初期化。$j = 0, l = 0$

STEP2 j 番目の階層において、クラスタ l に属する K 個の HMnet $\{m_1^l(j+1), \ldots, m_K^l(j+1)\}$ を用いて最大尤度を出力するクラスタを以下の式により選択。

$$\hat{l} = \underset{k}{\operatorname{argmax}} \sum_{i=1}^{I} L(m_k^l(j+1), a_i) \quad (1)$$

ここで $A = \{a_1, \ldots, a_I\}$ は入力データ、$L(m, a)$ は入力 a に対してモデル m から得られる対数尤度。

STEP3 対数尤度をバッファに記憶。尤度計算は以下の 2 種類について実験した。

(type1)

$$buf(j) = \max_{k} \sum_{i=1}^{I} L(m_k^l(j+1), a_i) \quad (2)$$

(type2) K 個の HMnet $\{m_1^l(j+1), \ldots, m_K^l(j+1)\}$ から、後述の話者混合法を用いて作成した K 混合の HMnet を $M^l(j+1)$ とする。

$$buf(j) = \sum_{i=1}^{I} L(M^l(j+1), a_i) \quad (3)$$

STEP4 $l \leftarrow \hat{l}, j \leftarrow j+1$

STEP5 $S_l(j)$ に対するサブクラスタが存在しない場合 STEP6 へ。それ以外は STEP2 に戻る。

STEP6 $buf(j)$ の値が最大となる階層、クラスタに属する HMnet を選択し認識に用いる。

2.4 認識アルゴリズム

クラスタリング木の選択されたノードにおける認識法として、以前提案している話者混合法を用いる [13]。話者混合法では、複数の同一構造を持つ HMnet の、対応する状態に含まれるガウス分布に等しい混合重みを与え、1 つの混合出力分布として表現することにより、混合連続出力分布 HMnet を作成する。本研究の場合は、各ノードごとに、K 個の HMnet を統合して K 混合の HMnet を作成し認識に用いる。

3 距離尺度

話者をクラスタリングするにあたって距離尺度が必要となる。比較的よくもちいられるものに、特徴量間の相関係数をその要素とする相関係数行列、ダイバージェンス (divergence)、バタチャリアの距離 (Bhattacharrya distance)、相互情報量などがある。これらの距離尺度

は、それぞれ異なった特徴をもっており、使われ方も異なっている。以下にそれぞれの特徴について説明する。

3.1 相互情報量

基本は、特徴量集合 $\mathcal{X}$ の任意の k 個の部分集合 $\mathcal{X}_k$ とクラスの集合 $\mathcal{C}$ 中の任意の l 個の組 $\mathcal{C}_l$ に対して定義できる相互情報量 $I(\mathcal{X}_k;\mathcal{C}_l)$ にある。これを必要に応じて単独に、あるいは適当に平均して用いる。数式的に計算できるのは正規分布や2値パターンなど限られた場合にとどまる。

3.2 相関係数行列

原則としてクラス間の分離に関する情報は持たない。例えば、既存の特徴量のおのおのとの相関を考慮して、なるべくいずれとも相関の高くない特徴量はそれだけ新しい情報をもたらす可能性が高いとみなす。すなわち、そのような特徴量を加えていけば新しい情報を有効に追加することができるのではないかと考えられる。

3.3 Bhattacharyya Distance

バタチャリアの距離 (Bhattacharyya distance) がある。バタチャリアの距離は一般的には、

$$D_B(f_1,f_2) = -\ln \int_{-\infty}^{\infty} \{f_1(x)f_2(x)\}^{1/2} dx \qquad (4)$$

これは、いわば2つの確率密度関数 $f_1(x)$ と $f_2(x)$ の間での相違の度合、あるいは分布間の距離のような尺度の一種であるとみなされる。

モデルの分布が正規分布で与えられるような場合、これを2クラスの分類の場合に限って話しをする。これは、2つのモデルの平均、分散をそれぞれ $\mu_1,\mu_2,\sigma_1,\sigma_2$ としたときに次式で与えられる。

$$\begin{aligned} D_B &= \frac{1}{8}(\mu_1-\mu_2)^t\left(\frac{\sigma_1+\sigma_2}{2}\right)^{-1}(\mu_1+\mu_2) \\ &+ \frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1|\cdot|\Sigma_2|}} \end{aligned} \qquad (5)$$

3.4 Kullback-Leibler に基づく距離尺度

Kullback-Leibler の情報量は

$$D_K = \sum_Y P(Y|M_1)\log\frac{P(Y|M_1)}{Y(N|M_2)} \qquad (6)$$

実際にはこの式は無限のデータを必要とする。実験では $P(Y|M_1)=1$ として、

$$D_K = \frac{1}{N_1}\sum_{k=1}^{N_1}\log\frac{P(Y_{1_k}|M_1)}{P(Y_{1_k}|M_2)} \qquad (7)$$

を用いた。以後この式を簡略化のために単に KL Distance あるいは KL 距離とよぶことにする。この距離尺度は明らかに非対称であるから、実際の計算においては対称化する必要がある。

これらのうち本研究では、Bhattacharyya 距離、KL 距離を用いることにする。

4 話者重み学習の検討

本研究で用いているアルゴリズムでは HMnet を用いており、各状態の出力確率密度はK混合ガウス分布に従うとしている。

この各分布は単一の話者のモデルから得られたもので、これを単純に組み合わせてモデルを構成し、認識を行なう。

このとき、単純 Baum-Welch 法により再学習をおこなったのでは、各状態の重み係数（＝話者の重み）が、ばらばらに学習されてしまう。

そこで各状態で重み係数を共通にすることで、選択された話者の重みによるモデルが作成され、より緻密なモデル化が期待できる。

5 実験

これらの理論を実際に確かめるために、次のような音素認識実験を行なった。

5.1 実験条件

音声資料には ATR データベースを用いた (表 5.1)。分析条件は表 4に示すとおりである。

表 4: 音響分析条件

音響特徴量	log パワー ＋16次 LPC ケプストラム ＋16次Δ LPC ケプストラム ＋Δ log パワー
サンプリング	12kHz
周期	5ms

表 5: 学習・評価データ

学習用データ			
話者	男性 85 人＋女性 85 人	サンプル数	50 文
適応用及び認識用データ			
話者	男性 9 人＋女性 9 人		
適応用	256 文節 (SB1 タスク) 先頭から N 文節 (N=1,3,5,7)		
認識用	279 文節 (SB3 タスク) 中の音素		

各話者ごとに単一ガウス分布２００状態 HMnet を作成した。ただし、HMnet の学習には大量の計算時間を必要とするため、実際には VFS を用いて話者適応を

している。このようにして作成された１７０のHMnetを$K=4$とおいて階層的にクラスタリングした。

5.2 クラスタリングの距離尺度に関する実験

上述したBhattacharyya距離とKLに基づく距離尺度を用いて話者をクラスタリングし、比較実験を行なった。予備実験として話者を男性話者５人＋女性話者５人で行ない、距離の計算に用いる文節の数を変えながら実験を行なった。実験結果を図3に示す。

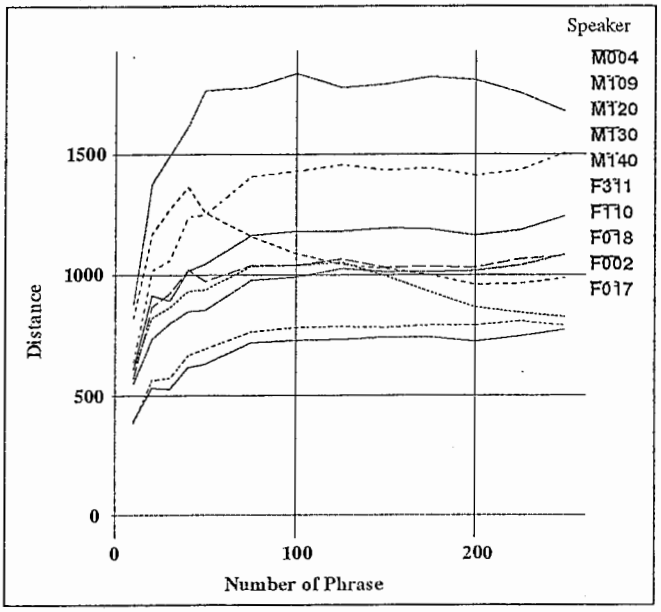

図3: KL距離の収束の様子

図3より文節の長さは１５０程度で十分だと思われるので、今後は１５０に固定して行なうことにした。

次にこの１０人を４つのクラスタの分類してみた。その結果を図5.2に示す。

そして、Bhattacharyya距離とKL距離による認識性能を比較した(表5.2)。

表6: 距離尺度と認識率(%)

	KL	Bhattacharyya
18人平均	69.52	69.17
FMSのみ	69.33	67.33

表5.2をみると全体ではKL距離の方が若干よい認識率であるが、話者によっては逆転する場合もある。有意な差は結局みられなかった。

5.3 話者結びつき重み学習に関する実験

話者結び重み学習に関する実験結果を表5.3に示す。

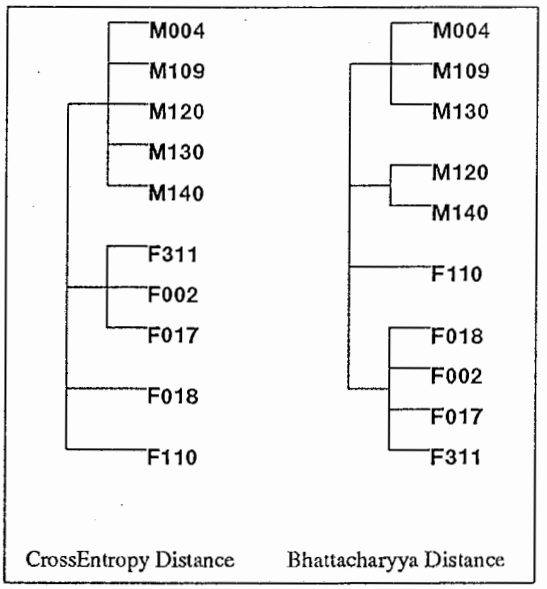

図4: 話者１０人のクラスタリングの結果

表7: 話者結び重み学習に関する実験(%)

適応に用いた文節数	話者結び重み	等重み
1	71.26	72.03
3	74.50	74.49
5	74.61	74.54
7	74.51	74.33

表5.3よりわかるように、有意義な差はみられなかった。実際に、学習された重みのパラメータをみてもあまり大きな変化はなかった。今回の実験の場合、混合数は4であるが、これをもっとふやせば有意な結果が得られる思われる。

6 まとめ

話者適応を用いた音声認識手法として、木構造を用いた話者クラスタリングアルゴリズムを提案しその有効性を音素認識実験により示した。

また、話者をクラスタリングするさいの距離尺度として、Bhattacharyya距離を用いて、よく用いられるKL距離尺度と比較実験をおこなった。その結果Bhattacharyya距離は、ほぼKL距離と同程度の分類能力をもつことがわかった。

さらに、適応話者が選ばれた後の、話者モデルの混合について話者結び重み法を提案し、実験をおこなったが、有意な差はみられなかった。

今後の展開としては以下のようなものがあげられよう。

1. 本アルゴリズムを話者の予備選択に用いて、VFSと組み合わせて話者適応を行なう

2. さらに話者数を増やして実験を行なう

謝辞
研究の機会を与えていただいた早稲田大学理工学部電気工学科白井克彦教授、小林哲則助教授、ATR 音声翻訳通信研究所社長山崎泰弘社長に感謝いたします。また、日頃から御指導いただく匂坂芳典室長、並びに熱心に討論いただく ATR のみなさまに感謝致します。

参考文献

[1] Paolo D'Orta, Marco Ferretti and Stefano Scarci, "Phoneme Classification for Real Time Speech Recognition of Italian", icassp, pp.81-84, 1987

[2] 相川 清明, 鹿野 清宏, 杉山 雅英, "距離最小化に基づく単語マルチテンプレート抽出法", 音響学会講演論文集, pp.115-116, 1982

[3] B.-H Juang and L. R. Rabiner, "Probabilistic Distance Measure for Hidden Markov Models", Bell System Technical Journal, 64, 1985

[4] 小坂 哲夫, 松永 昭一, 嵯峨山 茂樹, "木構造話者クラスタリングを用いた話者適応", 音響学会講演論文集, 1993

[5] 大倉 計美, 杉山 雅英, 嵯峨山 茂樹, "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", 電子情報通信学会技術研究会報告書, pp.23-28, SP92-16,1992,

[6] 小坂 哲夫, 鷹見 淳一, 嵯峨山 茂樹, "話者混合 SSS による不特定話者音声認識と話者適応", 電子情報通信学会技術研究会報告書, pp.17-24, SP92-52,1992.

[7] 坂元 慶行, 石黒 真木夫, 北川 源四郎, 情報量統計学, 共立出版株式会社

[8] 鳥脇 純一郎, 認識工学, コロナ社

[9] 杉山, 鹿野: "母音標準パタンの教師なし学習法," 音声研資, S83-48, 1983.

[10] 小坂, 牧野, 城戸: "男女声自動識別を用いた母音認識の検討," 音学講論, 1-9-12, 1984.

[11] 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] 菅村, 相川, 鹿野, 好田: "SPLIT, マルチテンプレート法による不特定話者単語音声認識," 音声研資, S82-64, 1989.

[13] 小坂, 鷹見, 嵯峨山: "話者混合 SSS による不特定話者音声認識と話者適応," 信学技報, SP92-52, 1992.

研究に使用したプログラムの説明

- ~/SSS/evaluate.csh_F_speaker10 話者と文節の長さを変えながら尤度を計算し lh_mat_file.(文節の長さ) というファイルに出力する。
- /SSS/calc_d3.csh というファイルにクラスタリング情報を書き込むシェル。出力フォーマットは、

```
10
300
9.307103e+05
7.770208e+05
7.667620e+05
7.970618e+05
```

のようになっている。上から話者の数、文節の長さ、尤度の順になっている。

- /SSS/calc_d3.c 標準入力から、話者の数、文節の数、距離のデータを読み込み平均をとり、標準出力に出力する C プログラム。出力フォーマットは、

```
# Speaker : (話者の数)
# MAXCOUNT : (文節の数)
# Average : (平均の距離)
```

となっている。

- /SSS/d3tograph.csh 話者と文節の長さを変えながら尤度を計算し lh_mat_file.(文節の長さ) というファイルに出力する。
- /SSS/d3tograph.c gnuplot ただし、予備実験なので女性話者 10 人で行なっている。
- /SSS/Exe_make_tree/link-zero.csh クラスタリングレベル 0(話者適応なし) のモデルにリンクをはるシェル。
- /SSS/Exe_make_tree/Reco_HMnet_main.csh
- /SSS/Exe_make_tree/Reco_HMnet.csh
- /SSS/Exe_make_tree/adapt_main.csh /SSS/Exe_make_tree/adapt_main.csh を用いて、話者を変えながら ADAPT(話者数)/adapt.(話者).(適応に用いる文節数) というファイルにクラスタリング情報を書き込むシェル。出力フォーマットは、

```
# 9494.162000 F306
# 9087.259000 F125
# 8578.618000 M126
# 8214.650000 M110
## 9378.573000
F306 1   F306 F007 F211 F010
```

のようになっている。

- /SSS/Exe_make_tree/adapt2.csh /SSS/Exe_make_tree/adapt_main.csh から呼ばれるシェル。
- /SSS/Exe_make_tree/spmix2.csh /SSS/Exe_make_tree/adapt2.csh から呼ばれるシェル。2 人の話者の HMnet のパラメータを Speaker Mixture 法を用いて新たに HMnet を構成する。
- /SSS/Exe_make_tree/spmix3.csh /SSS/Exe_make_tree/adapt2.csh から呼ばれるシェル。
- /SSS/Exe_make_tree/etof.c 標準入力から数字を %le フォーマットで読み込み、%f に変換して標準出力に出力する C プログラム

- /SSS/Exe_make_tree/Result170/get_rate_main.csh /SSS/Exe_make_tree/Result170/get_rate.csh を用いて、話者を変えながら認識率を result.(話者) というファイルに書き込むシェル。
- /SSS/Exe_make_tree/Result170/get_rate.csh 第一引数として話者を与えると、SSS-ToolKit/Tool/cal_result_26phone を用いて、第5位までの累積認識率を標準出力に出力するシェル。出力フォーマットは、

```
FAF/EVA0.1   72.23| 82.27| 88.00| 90.72| 92.67|
FAF/EVA1.1   75.51| 84.93| 91.14| 93.44| 94.98|
FAF/EVA1.2   72.85| 81.93| 87.37| 90.65| 92.74|
FAF/EVA1.3   72.51| 83.53| 88.83| 91.21| 93.16|
```

のように (話者)/EVA(適応に用いた文節の長さ).(クラスタリングレベル) (第一位の認識率)|(第二位までの累積認識率)|... となっている。

- /SSS/Exe_make_tree/rcg-rate-type1.csh /SSS/Exe_make_tree/Result170/get_rate.csh によってえられた結果から、type1 に基づいて尤度計算を行なう。
- /SSS/Exe_make_tree/rcg-rate-type2.csh /SSS/Exe_make_tree/Result170/get_rate.csh によってえられた結果から、type2 に基づいて尤度計算を行なう。
- /SSS/Src_clust/clust 話者クラスタリングプログラム