TR-IT-0045

話者クラスタリングを利用した
不特定話者音素モデルの作成法

倉岡 幹雄　小坂 哲夫
Mikio KURAOKA, Tetsuo KOSAKA

概要

従来、不特定話者音声認識を行なう場合、学習話者に対する検討が十分になされたとは言えない。特定話者音声認識におけるモデル作成法と同様な方法で、話者数だけを増加させ不特定話者モデルを作成する場合が普通であった。そこで本稿では、不特定話者音素モデル作成における話者の種類と話者数について検討を行なった。また話者数が増加した場合モデル作成に計算コストがかかるという問題があった。この問題に対し、パラメータの再推定をすることなく、特定話者音素モデルを合成し不特定話者音素モデルを作成する方法を提案する。以上について不特定話者音素認識実験により有効性を検討した。

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。本研究では、不特定話者認識の際、学習話者の選択方法と、特定話者の音声認識モデルを合成することにより、不特定話者の音声認識モデルを得る合成法について検討する。

近年において、認識タスクが高度化するにつれて、隠れマルコフモデル(Hidden Markov Model,HMM)で代表される確率的モデルが、重要な手法となっている。本実験では、逐次状態分割法(Successive State Splitting,SSS)により生成された、HMMの1つである隠れマルコフ網(Hidden Markov Network,HMnet)のモデルを使用した(図1)。

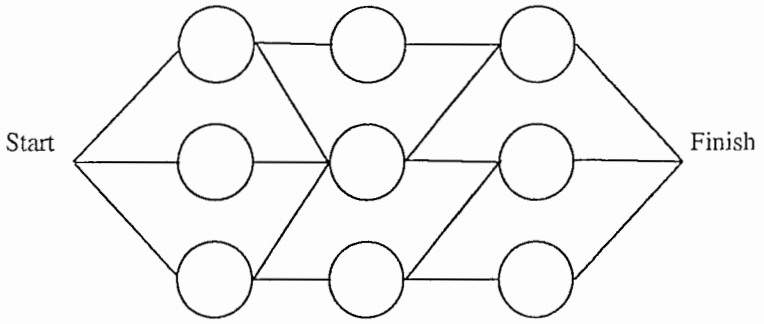

図 1: 隠れマルコフ網（HMnet）の例

不特定話者認識の学習法として一般に用いられている方法（Baum-Welchアルゴリズム）は、学習時間が非常にかかる。学習話者数が増えると学習時間も増えるため、学習話者数は上限があり、その上限に合わせるため学習話者を選択する必要がある。従来は無秩序に学習話者を選択していたが、どの学習話者が選ばれるかによって、不特定話者の認識率が変動するので、なるべく高い不特定話者の認識率を得る選択方法を考える必要がある。本研究では、モデル間の距離をBhattacharyya distanceを使用して定義し、各話者ごとに学習したモデルに対しクラスタリングを行ない、各クラスの典型的な話者を学習話者とする方法を行なった。この方法は、無秩序に選択した場合に比べて学習話者が２８５人から１０人選択する場合、不特定話者の認識率が向上した。

学習話者数と不特定認識率の関係を求める際、不特定学習話者数を変化させるとその学習話者数で再び学習をやり直さなければならない。これでは実験に時間がかかるため、学習の代わりに各話者ごとに学習したモデルを合成して不特定話者用のモデルを作った。

2 HMnetモデル間の距離

学習話者をクラスタリングするため、HMnetのモデル間の距離決めなければならない。距離を計る2つのモデルの構造は一致させ、各状態は混合数1の正規分布に従うとする。各状態は遷移確率と出力確率を持っているが、本実験では出力確率のみ、対応する状態間(図1の矢印で結ばれているもの)の距離を、次の式、Bhattacharyya distanceを用いて求める。

$$d(b^{(1)},b^{(2)}) = \frac{1}{8}(\mu_1-\mu_2)^t(\frac{\sum_1+\sum_2}{2})^{-1}(\mu_1-\mu_2) + \frac{1}{2}\ln\frac{|(\sum_1+\sum_2)/2|}{|\sum_1|^{1/2}+|\sum_2|^{1/2}} \quad (1)$$

ただし、
$b^{(i)}$: 正規分布のモデルiの出力確率
μ_i: モデルiの状態の出力確率の期待値
$\sum_i$: モデルiの出力確率の状態の分散
すべての状態で$d(b^{(1)},b^{(2)})$を求め、それを合計したものを2つのモデル間の距離とする。

3 クラスタリング

各話者ごとに学習を行ない、それぞれ混合数1のモデルがある。その話者の母集団をBhattacharyya distanceを用いてクラスタリングを行なう。本実験で使用したクラスタリングの方法を解説する。

1 話者N人のモデル計算。（図3の１）

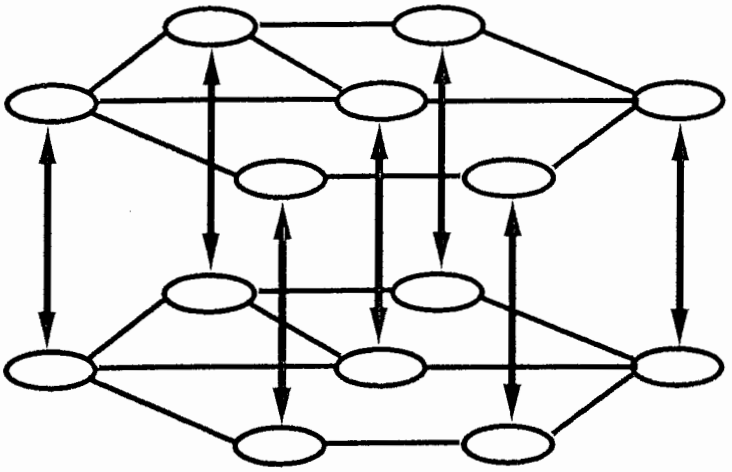

図 2: モデル間の距離の例

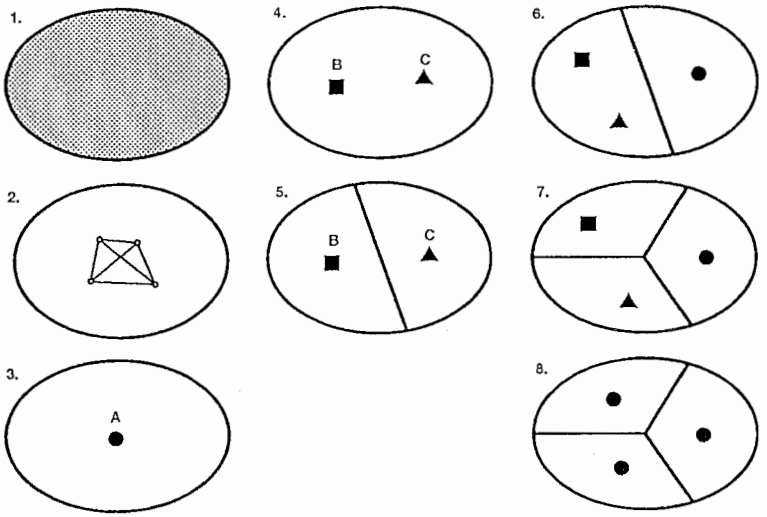

図 3: クラスタリング

2 $N \times N$ の話者間距離のマトリックスを作成。（図 3の 2）

3 各話者からの距離の合計が最小となる話者 A を選択。（図 3の 3）

4 話者 B、C を選択。条件として、各話者から BC のうち近い方への距離の合計が最小となるようにする。（図 3の 4）

5 すべてのグループのグループ内の各話者と近い中心核の距離の和を D_{AV}、閾値を TH とする。$D_{AV} < TH$ ならば、終了。

6 グループ分け。各話者をもっとも近い中心核（例： B or C）に属させる。（図 3の 5）

7 中心核への距離の和がもっとも大きいグループについて新たに 2 つの中心核を選択。条件 4 と同様。（図 3の 6）

8 グループ分け（再配分）。条件 6 と同様。（図 3の 7）
$i = 1$

9 各グループについて各話者からの距離の平均が最小になる中心核を選ぶ。（図 3の 8）
$i = i + 1$

10 $D_{AVi} = D_{AV(i-1)}$ ならば条件 11 へ、違うなら条件 8 へ。

11 $D_{AV} < TH$ ならば終了、ちがうならば条件 4 へ。

4 HMnet 合成の方法

学習話者を変化させた場合、学習を始めからやりなおさなければならない。学習話者と認識率の関係を調べる場合、これでは学習時間がかかってしまうので、モデルを合成した。N 人の学習話者から学習話者 m 人、混合数 $k(< m)$ となるモデルの合成方法を示す。本実験で用いた合成前のモデルは、学習者 1 人で条件 1 を満たした混合数 1 のモデルである。合成方法を図 4 に示す。この方法は、次の 3 つの手順からなる。

1. 学習話者のクラスタリング
   学習話者をクラスタリングによりクラス $1, 2, ..., m$ に分ける。$n(i)$ はクラス i の人数（モデル数）である。

2. モデルの平均化（Average method）
   クラス内のモデル構造の同じ位置にある状態の遷移確率、出力確率の平均値、分散を平均化する。

3. モデルの多混合化（Speaker mixture method）
   混合数 1 の各クラスのモデルを多混合化することで混合数 m のモデルを作る。出力分布は、クラスのメンバー数に比例する場合と、クラスのメンバー数に関係なく一定の場合が考えられる。

4.1 正規分布で表された群の和

合成法の手順 2 で状態の平均値、分散を平均化する方法は、数群の平均化の式を用いた。
$\bar{X}$: 平均化された平均値
S: 平均化された偏差平方和
V: 平均化された分散 n_i : 話者 i のサンプル数
V_i : 話者 i の分散

$$S_i = (\bar{X}_i - \bar{X})^2 \tag{2}$$

平均値

$$\bar{X} = \frac{\sum n_i \bar{X}_i}{\sum n_i} \tag{3}$$

分散

$$V = \frac{V_i}{\sum n_i} + \sum \frac{n_i}{\sum n_i} (\bar{X}_i - \bar{X})^2 \tag{4}$$

本実験では、合成前のモデルの学習条件が同じためサンプル数は、話者によらず一定とした。

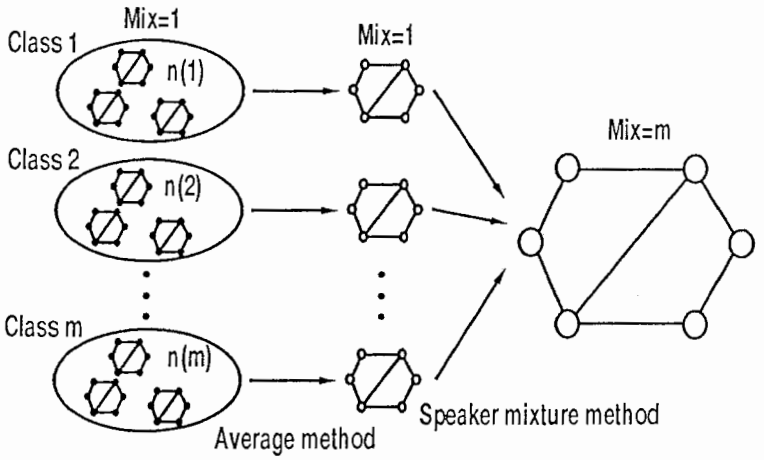

図 4: モデル合成

5 実験

学習条件を図 1 に示す。

表 1: 学習・評価データ

分析条件	
Sampling-rate	12kHz
Window	Hamming window (20ms)
Frame period	5ms
Analysis	log power + 16-order LPC-Cep + Δlog power + 16-order ΔLPC-Cep
学習用データ	
話者	285 人から N 人を選択
データ	C-Set 50 文
学習方法	VFS
HMnet	200 状態、単一ガウス分布, 対角共分散行列使用
認識用データ	
話者	6 人 (FAF,FKN,FMS,MHT,MMS,MMY)
データ	279 文節 (SB3 タスク) 中から切り出された音素

5.1 話者クラスタリング

図 5.1 に、２８５人から学習話者をランダムに選んだ場合と話者クラスタリングによって、選んだ場合の学習話者数と距離の比較を示す。この距離は、クラス中心のモデルと、クラス内のすべてのモデルとの距離を合計し、その距離をすべてのクラスについて求めて合計したもので、平均値ではない。ランダムに話者を選ぶのは１つの学習話者数につき５回行なった。図 5.1 において、randamize set が各学習話者数に渡っているが、互いに関係はない。このけっかよりクラスタリングは正しく行なわれており、ランダム選択の５人の場合の距離と、クラスタリング選択の５人の場合の距離がほぼ等しいことがわかる。

5.2 学習話者と不特定話者認識率

図 5.4 〜図 5.4 は学習話者と不特定話者認識率の関係を示したものである。まず、図 5.4 〜図 5.4 を比較すると、学習話者５人と１０人の場合で認識率に差が出ている。図 5.4 〜図 5.4 も同様の傾向である。次に図 5.4 と図 5.4 を

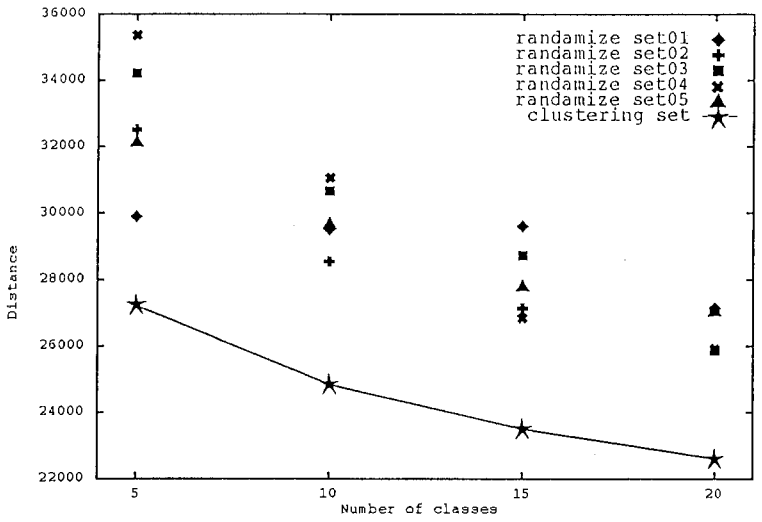

図 5: 学習話者数と距離

比較すると、クラスのメンバー数に比例する方が認識率がよい。

5.3 文節音声認識実験

合成法によりモデルを作成し、文節音声認識を行なった。学習条件を表 2 に示す。認識結果を表 3 に示す。比較のため音素の認識結果を表 4 に示す。

表 2: 文節音声認識実験条件

手法	SSS-LR 連続音声認識法
ビーム幅	1200
タスク	国際会議の参加予約に関する対話 (複合語を許さない文節発音)
文節内文法	
規則数	1407
語彙数	1035
音素 Perplexity	5.9
単語 Perplexity	約 100

表 3: 合成法により 285 を N 混合にした場合の不特定話者文節認識率

混合数	候補	FAF	FKN	FMS	MHT	MMS	MMY	平均値
5Mix.	1 位	80.65	76.62	77.70	87.68	80.58	74.10	79.55
	5 位	96.06	94.60	93.88	97.46	96.40	97.12	95.92
10Mix.	1 位	82.08	78.06	78.78	89.13	81.65	73.74	80.57
	5 位	96.42	95.32	94.24	98.19	96.04	97.12	96.22

5.4 クラスタリング選択 / ランダム選択の比較

表 4: 合成法により 285 人を N 混合にした場合の不特定話者音素認識率

混合数	候補	FAF	FKN	FMS	MHT	MMS	MMY	平均値
5Mix.	1 位	76.27	73.30	74.06	82.20	75.94	72.80	75.76
	5 位	95.32	96.80	94.92	98.89	97.49	96.98	96.73
10Mix.	1 位	77.11	75.10	75.10	82.75	76.29	73.30	76.60
	5 位	95.53	97.01	95.06	98.82	97.84	97.05	96.88

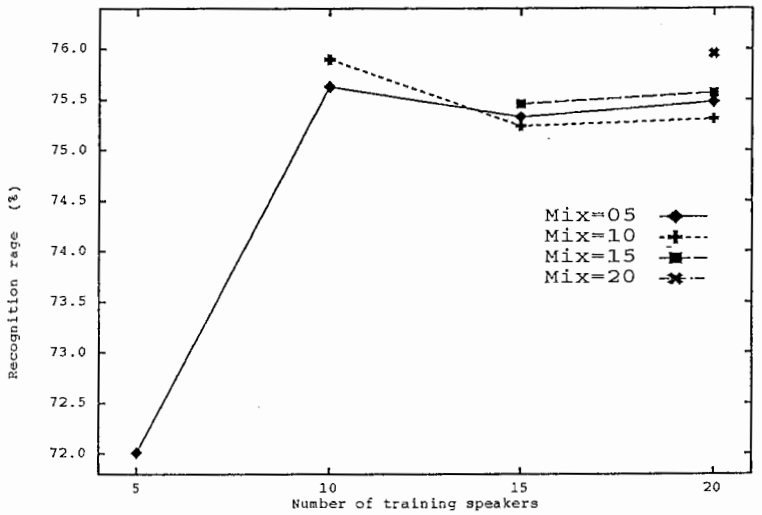

図 6: 学習話者と不特定話者認識率 (クラスタリング ; 出力確率は一定)

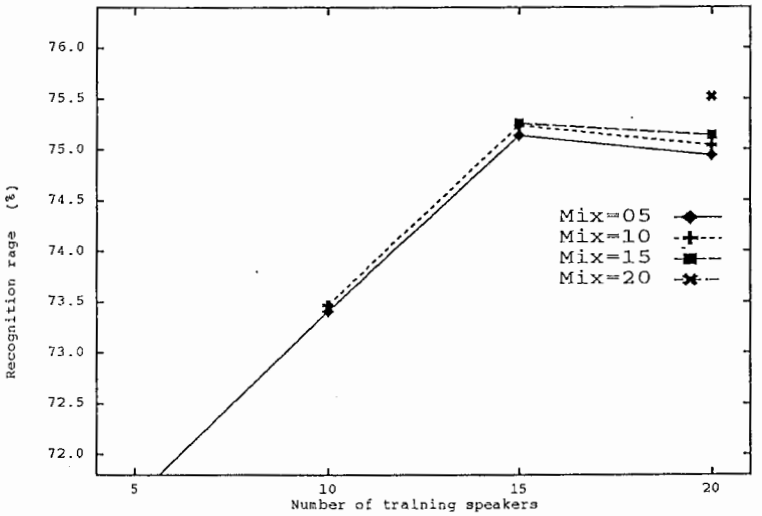

図 7: 学習話者と不特定話者認識率 (ランダム 5 回の平均 ; 出力確率は一定)

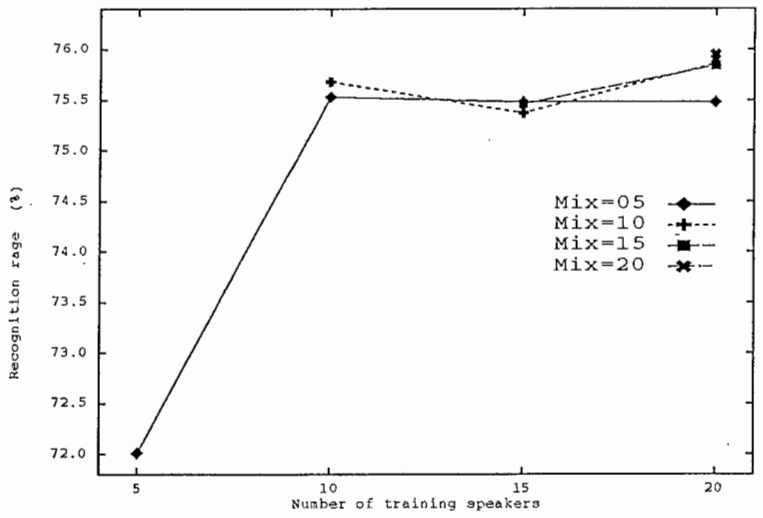

図 8: 学習話者と不特定話者認識率 (クラスタリング；出力確率はクラス数に比例)

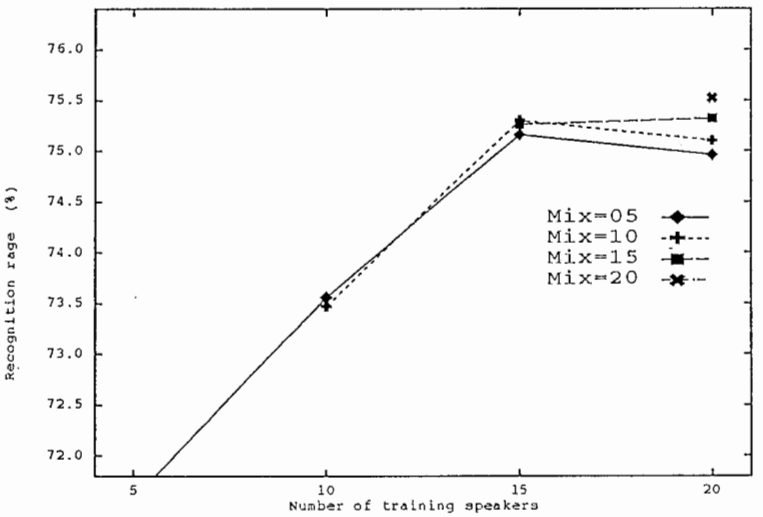

図 9: 学習話者と不特定話者認識率学習話者と不特定話者認識率 (ランダム 5 回の平均；出力確率はクラス数に比例)

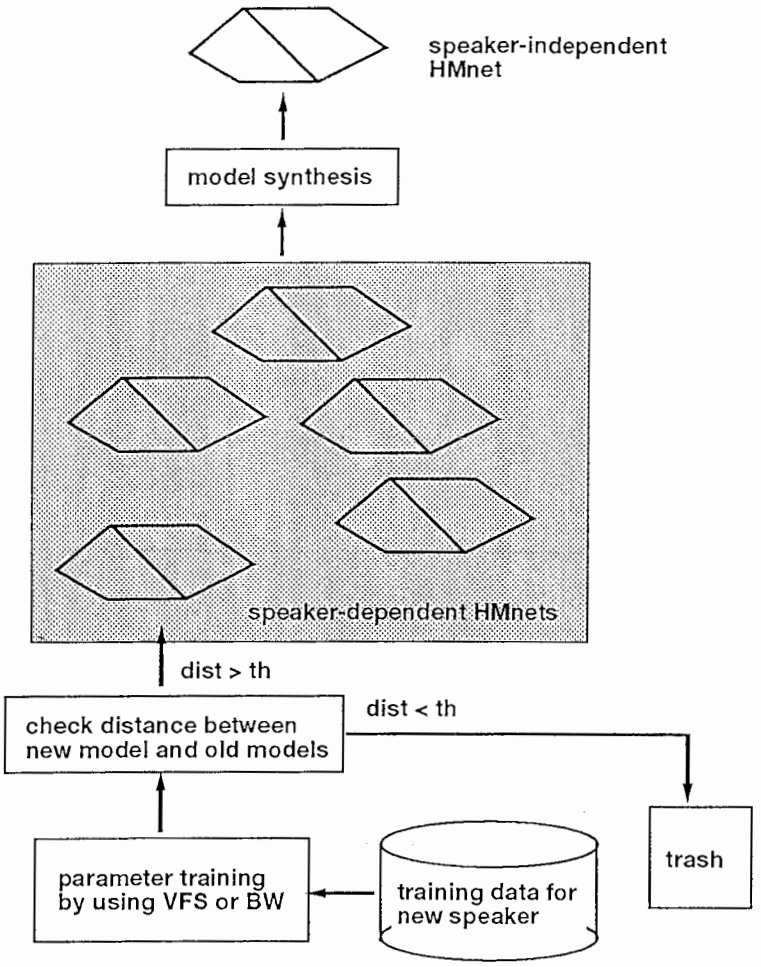

図 10: 学習話者数と距離

6 まとめ

これらの結果から以下のことがいえる。

・話者クラスタリングによる学習話者の選択方法は、学習話者が５人、１０人の少ない場合有効である。

7 今後の課題

・合成方法と他の学習方法との比較

・大多数学習話者による認識実験の検討

・モデル合成法による不特定話者の認識の検討（図 7）

謝辞

研究の機会を与えていただいた豊橋技術科学大学情報工学系計算機大講座中川聖一教授に感謝いたします。また、御指導いただいた ATR のみなさまに感謝いたします。

8 プログラムについて

この実験に使用したプログラムの説明と使用方法を示す。

8.1 プログラム説明

~/SSS/
  Average/ 出力確率が1／mixの場合の
    Member_list.rand/ 学習メンバーリストのディレクトリ
  Hirei_Average/ 出力確率がクラスのメンバー数に比例
  Data/ 認識用不特定話者音声データ
  tool/ 学習・認識ツール群

~/SSS/tool/
  Average_HMnet.csh 出力確率が1／mixのモデル合成を行なう
  Hirei_Average_HMnet.csh 出力確率がクラスのメンバー数に比例するモデル合成を行なう
  Reco_HMnet.csh 不特定話者6人による認識
  Adapt_HMnetA.csh 50文のVFSによる話者適用
  Adapt_HMnet.csh 150文のVFSによる話者適用
  Retrain_HMnet-no_init.csh モデルの初期化を行なわずに再学習
  get_rate.csh 認識率計算（学習話者別）
  get_rate_mix.csh 認識率計算（平均値）

~/Src_clust/
  clust-sil.csh 話者クラスタリング（学習話者の順番を出力）
  randomize-sil.csh ランダムに学習話者を選択（学習話者の順番を出力）
  belong.csh 学習話者の順番を学習話者名のリストに変換
  make_condition.csh 学習話者名のリストからファイルのリストを制作
  make_member.csh 学習話者名のリストから学習話者メンバーリストを制作

8.2 使用方法

・285人からクラスタリングにより学習話者数20、混合数15のメンバーリストを作る場合。

```
cd ~/Src-clust
clust-sil.csh 20 condition285 > temp1
belong.csh temp1 condition285 > temp2
make_condition.csh temp2 condition285 > condition.clust.me020
clust-sil.csh 15 condition.clust.me020 > temp3
belong.csh temp3 condition.clust.me020 > temp4
make_member.csh temp4 > member.clust.me020mx15
cp member.clust.me020mx15 ~/SSS/Average/Member_list
```

・285人からランダムに学習話者数20、混合数15のメンバーリストを作る場合。

```
cd ~/Src-clust
randomize-sil.csh 20 condition285 > temp1
belong.csh temp1 condition285 > temp2
make_condition.csh temp2 condition285 > condition.rand.me020
rand-sil.csh 15 condition.rand.me020 > temp3
belong.csh temp3 condition.rand.me020 > temp4
make_member.csh temp4 > member.rand.me020mx15
cp member.rand.me020mx15 ~/SSS/Average/Member_list
```

・member.clust.me020mx15 から合成方法 (Average) によりモデルを作る場合。

```
cd ~/SSS/Average
Average_HMnet.csh Member_list/member.clust.me020mx15
                  model.clust.me020mx15.ave
```

・member.clust.me020mx15 から合成方法 (Hirei_Average) によりモデルを作る場合。

```
cd ~/SSS/Hirei_Average
Hirei_Average_HMnet.csh Member_list/member.clust.me020mx15
                        model.clust.me020mx15.h_ave
```

・model.clust.me020mx15.ave から VFS により話者適応モデルを作る場合。

```
cd ~/SSS/Average
Adapt_HMnetA.csh model.clust.me020mx15.ave
                 model.clust.me020mx15.ave.vfs
```

・model.clust.me020mx15.ave から Retrain によりモデルを作る場合。

```
cd ~/SSS/Average
Retrain_HMnet-no_init.csh model.clust.me020mx15.ave
                          model.clust.me020mx15.ave.tra
```