TR-I-0355

不特定話者混合連続分布型 HMnet 作成ユーザーズマニュアル
Continuous Mixture HMnet for Speaker-Independent Speech Recognition User's Manual

小坂 哲夫　鷹見 淳一　嵯峨山 茂樹
Tetsuo KOSAKA, Jun-Ichi Takami
and Shigeki SAGAYAMA

概要

本レポートでは不特定話者用混合連続分布型 HMnet(Hidden Markov Network) の音素モデル作成を行なうプログラムについて、その使用方法を説明する。また、Cシェルスクリプトで書かれた実際の使用例も添付する。本レポートで扱う範囲は音素モデルの作成のみで音声パラメータファイルの作成部及び、認識システムは含まない。

目次

第 1 章

はじめに

混合連続分布型 HMnet(CM-HMnet) の音素モデルの作成を行なうプログラムの使用方法について説明する。 HMnet は状態を音素間で共有することを特徴とした音素環境依存モデルであり [1]、本パッケージは話者混合法を用いて不特定話者用 HMnet を作成するものである。さらに本パッケージには話者重み学習による話者適応プログラム、話者プルーニング法によるモデル混合数削減プログラムも含まれている。また SSS-LR プログラムを購入した場合、不特定話者認識システムとして使えるよう不特定話者音素モデルも含まれている。

話者混合法、話者重み学習法、話者プルーニング法については、文献 [3] を参考のこと。

本パッケージに含まれる内容について以下に示す。

【ディレクトリーの構成】

```
|
+- Data 音声パラメータファイル・ラベルファイル
|   |
|   +- M001 話者 M001
|   +- M002 話者 M002
|   +- M003 話者 M003
|
+- Etc その他
+- Exe 実行ファイル格納ディレクトリー
+- Inmatrix イニシャル HMnet
+- Model 不特定話者 HMnet
|   |
|   +- mix12f 女性話者用 HMnet
|   +- mix12m 男性話者用 HMnet
|   +- mix12mf 男女話者用 HMnet
|
+- Outmatrix 作成ファイル格納用ディレクトリ
|   |
|   +- M001 話者 M001 用 HMnet
|   +- M002 話者 M002 用 HMnet
|   +- mix2 混合数 2 用 HMnet
|
+- Src ソースファイル
```

【ファイルの内容】
以下にシェルスクリプトの内容を示す。

adapt_VFS.csh	VFS 法による話者適応シェルスクリプト
adapt_STWT.csh	話者重み学習による話者適応シェルスクリプト
lab_conv.csh	ラベル変換用シェルスクリプト
example.csh	CM-HMnet 作成サンプルシェルスクリプト

以下の実行ファイルは make 実行後に作成される。 make については「サンプルプログラムの実行例」の章を参考のこと。

Exe/Adapt	VFS 法による特定話者 HMnet 作成プログラム
Exe/Cut_sp	スピーカー・プルーニング法による混合数削減プログラム
Exe/Lab_conv	ラベル変換プログラム
Exe/Make_list	音素ラベル抽出プログラム
Exe/Make_std_form_from_ascii	モデル変換プログラム 1
Exe/Mksmsss	CM-HMnet 作成プログラム
Exe/Model_conv	モデル変換プログラム 2
Exe/Train_weight_tied	話者重み学習による話者適応プログラム

また、メディアに含まれる全てのファイルのリストを本文末に掲載した。

この章の残りでは CM-HMnet ソフトウェアパッケージの使用上の制限について述べる。第2章ではデータやファイルのフォーマット、全プログラムの使い方について述べる。第3章ではインストールについて述べるとともに、混合数が 2 の HMnet 作成の実行例も示す。

1.1 使用上の注意

このソフトウェアは HP9000/730(OS は HP-UX 8.05) 上でコンパイルされ、動作が確認されている。必要とされるメインメモリは学習データ量に依存する。動作を確認した我々のワークステーションは 64MB のメモリを実装している。添付してあるデモソフトを実行するには、 csh などの UNIX プログラムが必要である。またプログラムソースはディレクトリ Src にある。

第 2 章

プログラムの説明

2.1 不特定話者用 CM-HMnet

本パッケージには、別売の SSS-LR プログラムと組み合わせて、不特定話者音声認識のシステムが組めるよう、男性話者用・女性話者用・男女話者用の音素モデル (CM-HMnet) が含まれている。音素モデルはアスキーテキストであり more などで確認できる。

```
Model/mix12mf/matrix.200        # 男女話者用、状態数 200
Model/mix12mf/matrix.600        # 男女話者用、状態数 600
Model/mix12m/matrix.200         # 男性話者用、状態数 200
Model/mix12m/matrix.600         # 男性話者用、状態数 600
Model/mix12f/matrix.200         # 女性話者用、状態数 200
Model/mix12f/matrix.600         # 女性話者用、状態数 600
```

SSS-LR を不特定話者で動かす場合の引数の設定例を以下に示す。

```
-y 2.0          # Sigma_Scale
-N 2.4          # n*state-standard-deviation
-K 1            # use of multimodel
-P 5            # frame-period (msec)
-B 256          # global-beam
-b 64           # local-beam
-a 4.0          # dur conv break point
-R 0.6          # dur conv rate
-c 4000         # total-cells
-A 15.0         # add-time (msec)
```

引数の詳細な説明については、別売の SSS-LR ソフトを参照のこと。

2.2 音響パラメータファイルと音素ラベルファイル

音響パラメータと音素ラベルはそれぞれ別のファイルとなっている。音響パラメータとしては、図 2.1に示すように、通常、 log power, cepstrum(16 次), Δlog power, Δcepstrum(16 次) から成る 4 種類のファイルを用いている。ラベルファイルは各音素ラベルの開始フレーム、継続フレームおよび前後の音素環境を記述したものである。

本パッケージではデモ用として、 ATR データベース Cset(不特定話者セット) の B タスク (音素バランス 216 単語) の 3 名分についてパラメータ及び、ラベルファイルを用意した。

log power, cepstrum(16 次), Δlog power, Δcepstrum(16 次) から成る 34 次元ベクトルの作成方法については別売の分析ツールを参照のこと。

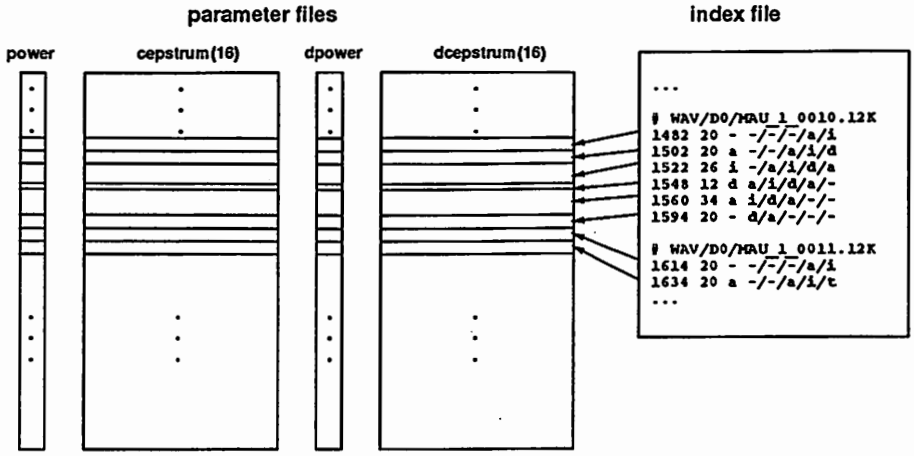

図 2.1: 音響パラメータとラベル（インデックス）ファイルの関係

話者 M001 の音素バランス 216 単語は以下の 4 つのファイルである。

```
Data/M001/M001_B.5mS.lpow      # ログパワー
Data/M001/M001_B.5mS.cep16     # ケプストラム (16 次)
Data/M001/M001_B.5mS.dpow      # デルタログパワー
Data/M001/M001_B.5mS.dcep16    # デルタケプストラム (16 次)
```

以下のラベルファイルには、上の音響パラメータファイルの単語に対応する音素ラベル、開始フレーム、継続フレームおよび前後の音素環境が記述されている。

```
Data/M001/M001_B.5mS.cxt
```

他の話者 M002・M003 についても同様の構造になっている。

2.3 ラベルの変換

ATR データベースのラベルフォーマットと本プログラムで用いるラベルフォーマットは一部ことなるので、ラベルファイルを使用する前に、フォーマット変換を行なう必要がある。

このフォーマット変換には C-Shell プログラムの lab_conv.csh を使用する。

【lab_conv.csh の使用法】

```
lab_conv.csh ATR ラベルファイル名 出力ラベルファイル名
```

【lab_conv.csh の使用例】

```
lab_conv.csh Data/M001/M001_B.5mS.cxt Data/M001/M001_B.5mS_ng.cxt
```

2.4 不特定話者用 CM-HMnet の作成法

2.4.1 概要

不特定話者用 CM-HMnet 作成法の概要を示す。詳しくは文献 [3] を参照のこと。

学習は Baum-Welch アルゴリズムが用いられている。学習サンプルすべてをメモリに取り込むため、母音などのサンプル数の多い音素については大きなメモリ空間を必要とするので要注意。

1. 初期 HMnet の作成

   出力分布が単一ガウス分布であるような初期 HMnet を、話者 1 名の大量のデータから SSS アルゴリズムを用いて生成する。

2. パラメータ学習

   次に初期 HMnet に対し、 VFS 法を利用して複数の話者の比較的少量のデータによりそれぞれ適応をおこない、複数話者分の HMnet を作成する。

3. 話者混合化

   この複数話者分の HMnet の対応する状態を 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。このモデルにより不特定話者音声認識を行なう。

4. 話者重み学習による話者適応

   極少量の入力データから、話者重み学習により高速な話者適応を行なう。

5. 話者プルーニング

   話者重みの低い話者に対応する混合成分を削除することにより、認識時の計算量の削減を行なう。

2.4.2 初期 HMnet の作成

本パッケージには作成済の初期 HMnet が以下のように含まれている。

```
Inmatrix/matrix.200      # 状態数 200 の初期 HMnet
Inmatrix/matrix.300      # 状態数 300 の初期 HMnet
Inmatrix/matrix.400      # 状態数 400 の初期 HMnet
Inmatrix/matrix.500      # 状態数 500 の初期 HMnet
Inmatrix/matrix.600      # 状態数 600 の初期 HMnet
```

また同ディレクトリーにあるファイル名の先頭に phone とつくファイルは HMnet がカバーしない音素環境のための、汎用の音素モデルである。

2.4.3 パラメータ学習

次に初期 HMnet に対し、 VFS による話者適応法を利用して複数の話者の比較的少量のデータによりそれぞれ適応をおこない、複数話者分の HMnet を作成する。作成された HMnet は特定話者音素モデルとして利用できる。 HMnet の VFS による話者適応については文献 [2] を参照のこと。

VFS による話者適応は C-Shell プログラムの adapt_VFS.csh を使用する。

【adapt_VFS.csh の使用法】

adapt_VFS.csh 状態数 適応単語数 初期 HMnet ディレクトリー名
出力 HMnet ディレクトリー名 ログパワーファイル名 ケプストラムファイル名
Δ ログパワーファイル名 Δ ケプストラムファイル名 ラベルファイル名

以上により、出力 HMnet ディレクトリーに特定話者に適応後の HMnet が得られる。但し入・出力の HMnet のファイル名は"matrix.200" など"matrix."+ 状態数とする。

【adapt_VFS.csh の使用例】

```
adapt_VFS.csh  200 216 Inmatrix Outmatrix/M001\
           Data/M001/M001_B.5mS.lpow Data/M001/M001_B.5mS.cep16\
           Data/M001/M001_B.5mS.dpow Data/M001/M001_B.5mS.dcep16\
           Data/M001/M001_B.5mS_ng.cxt > Log/log_VFS_M001.txt
```

【adapt_VFS.csh の出力】
adapt_VFS.csh を実行すると以下の出力が stdout に得られる。

- 学習繰り返しごとの HMnet の対数尤度出力とその差分。
- 学習繰り返し回数。
- 最終的な HMnet 対数尤度出力。
- VFS 法におけるスムージング率及び、分散適応したかどうかのフラグ。

【adapt_VFS.csh の出力例】

```
{>  0  1.461220e+06  0.000000e+00}
{>  1  2.033885e+06  2.815623e-01}
{>  2  2.063331e+06  1.427090e-02}
{>  3  2.069551e+06  3.005374e-03}
{>  4  2.071997e+06  1.180783e-03}
{>  5  2.073501e+06  7.252702e-04}
{# iteration times : 6}
{# total prob      : 2.073980e+06}

smoothing_rate = 5 , var_adapt_sw = 0
```

2.4.4 話者混合化

話者混合法を用い、複数話者分の HMnet の対応する状態を 1 つの混合出力分布として表現して、混合連続出力分布 HMnet(CM-HMnet) を作成する。作成されたモデルは不特定話者音声認識用の HMnet として用いることができる。
話者混合による CM-HMnet 作成は C プログラムの Mksmsss を使用する。

【MKsmsss の使用法】

Exe/Mksmsss 入力 HMnet のリスト 出力 HMnet 名

以上により入力 HMnet のリストに記載された特定話者用 HMnet が混合化され 1 つの CM-HMnet が作成される。

入力 HMnet のリストのフォーマットは特定話者用 HMnet 名をディレクトリーも含めて書く。例を以下に記載する。

```
Outmatrix/M001/matrix.200
Outmatrix/M002/matrix.200
```

【MKsmsss の使用例】

```
mkdir -p Outmatrix/mix2
touch .list_smsss
echo "Outmatrix/M001/matrix.200" >> .list_smsss
echo "Outmatrix/M002/matrix.200" >> .list_smsss
Exe/Mksmsss .list_smsss Outmatrix/mix2/matrix.200
rm .list_smsss
```

2.4.5 話者重み学習による話者適応

極少量の入力データから、話者重み学習により高速な話者適応を行なうには C-Shell プログラムの adapt_STWT.csh を使用する。

【adapt_STWT.csh の使用法】

adapt_STWT.csh 適応語数 入力 HMnet 名 出力 HMnet 名
ログパワーファイル名 ケプストラムファイル名 Δ ログパワーファイル名
Δ ケプストラムファイル名 ラベルファイル名

以上によりラベルファイルの先頭から適応語数分のデータを用い、入力 HMnet の話者適応を行ない、出力 HMnet を生成する。

【adapt_STWT.csh の使用例】

```
adapt_STWT.csh 10 Outmatrix/mix2/matrix.200 Outmatrix/mix2/matrix.200.STWT\
Data/M003/M003_B.5mS.lpow Data/M003/M003_B.5mS.cep16\
Data/M003/M003_B.5mS.dpow Data/M003/M003_B.5mS.dcep16\
Data/M003/M003_B.5mS_ng.cxt > Log/log_STWT_M003.txt
```

【adapt_STWT.csh の出力】
adapt_STWT.csh を実行すると以下の出力が stdout に得られる。

- 学習繰り返しごとの HMnet の対数尤度出力とその差分。
- 学習繰り返し回数。
- 最終的な HMnet 対数尤度出力。

【adapt_VFS.csh の出力例】

```
{>  0  6.669417e+04  0.000000e+00}
{>  1  6.684344e+04  2.233171e-03}
{>  2  6.685565e+04  1.825659e-04}
{>  3  6.685718e+04  2.284378e-05}
{# iteration times : 4}
{# total prob      : 6.685741e+04}
```

2.4.6 話者プルーニング

話者重みの低い話者に対応する混合成分を削除することにより、認識時の計算量の削減を行なう。

話者プルーニング法による混合数削減 HMnet 作成には C プログラムの Cut_sp を用いる。

【Cut_sp の使用法】

Exe/Cut_sp 入力 HMnet 名 出力 HMnet 名 スレショルド

以上により入力 HMnet のうち、スレショルド以下の値の混合成分が除去され混合数が削減した出力 HMnet が生成される。

【Cut_sp の使用例】

```
Exe/Cut_sp Outmatrix/mix2/matrix.200.STWT\
Outmatrix/mix2/matrix.200.STWT.PR 0.27
```

2.4.7 モデルのフォーマット変換

別売の SSS-LR と組み合わせて音素モデルを用いる場合、このままではフォーマットが異なるので、フォーマット変換を行なう。フォーマット変換には C プログラムの Exe/Make_std_form_from_asci 及び、Exe/Model_conv を用いる。

【フォーマット変換プログラムの使用法】

Exe/Make_std_form_from_ascii 入力 HMnet 名 | Exe/Model_conv > 出力 HMnet 名

【フォーマット変換プログラムの使用例】

```
Exe/Make_std_form_from_ascii matrix.200 | Exe/Model_conv > matrix.200.NEW
```

第 3 章

サンプルプログラムの実行例

本パッケージには C-Shell プログラムで書かれた、サンプルプログラムが含まれている。以下にその実行方法について記す。実行するにはディスクにおよそ 30MB の空きスペースが必要である。

【実行内容】
サンプルプログラムを実行することにより、以下の事項が行なわれる。

1. ラベル変換
2. VFS による話者 M001 及び話者 M002 用特定話者音素モデルの作成
3. 話者混合法による 2 混合の HMnet の作成
4. 話者重み学習による話者 M003 への話者適応
5. 話者プルーニングによる 2 混合 HMnet の混合数削減
6. モデルのフォーマット変換

【実行方法】

```
$ mkdir CM-SSS      # ディレクトリー作成
$ cd CM-SSS         # ディレクトリーの変更
$ tar xvf           # メディアからのプログラムなどの読み込み
$ cd Src            # ディレクトリーの変更
$ make              # 実行ファイルの作成
                    # ディレクトリー''Exe'' に 6 つの実行ファイルが作成される
$ cd ..             # ディレクトリーの変更
$ example.csh       # サンプルプログラムの実行
```

【出力】
以下のファイルまたはディレクトリーが作成される。

```
Data/M001/M001_B.5mS_ng.cxt          # フォーマット変換されたラベルファイル
Data/M002/M002_B.5mS_ng.cxt          # フォーマット変換されたラベルファイル
Data/M003/M003_B.5mS_ng.cxt          # フォーマット変換されたラベルファイル
Outmatrix/M001/matrix.200.NEW        # VFS 法による話者 M001 に適応後の HMnet
Outmatrix/M002/matrix.200.NEW        # VFS 法による話者 M002 に適応後の HMnet
Log                                  # 出力ログ用ディレクトリー
```

```
Log/log_VFS_M001.txt                    # VFS 法による話者 M001 への適応実行時の
ログ
Log/log_VFS_M002.txt                    # VFS 法による話者 M002 への適応実行時の
ログ
Outmatrix/mix2/matrix.200.NEW           # 話者混合法による 2 混合 HMnet
Outmatrix/mix2/matrix.200.STWT.NEW      # 話者重み学習法による重み変更後の HMnet
Log/log_STWT_M003.txt                   # 話者重み学習時のログ
Outmatrix/mix2/matrix.200.STWT.PR.NEW# 話者プルーニングによる混合数削
                                        # 減後の HMnet
```

【動作の確認】

本パッケージには、HP9000/730 でサンプルプログラムの実行により作成された HMnet が含まれている。これと今回作成した HMnet を diff などの unix コマンドにより比較することにより、正常に動作したか確認できる。

但し、計算機が異なると誤差などにより混合係数や平均値の値などが完全には一致しないこともあるので注意のこと。

本パッケージに含まれる、サンプルプログラムにより作成された HMnet のリスト。

```
Outmatrix/M001/matrix.200.NEW.HP
Outmatrix/M002/matrix.200.NEW.HP
Outmatrix/mix2/matrix.200.NEW.HP
Outmatrix/mix2/matrix.200.STWT.NEW.HP
Outmatrix/mix2/matrix.200.STWT.PR.NEW.HP
```

全ファイルのリスト

```
Data/
Etc/
Exe/
Inmatrix/
Model/
Outmatrix/
README
Src/
adapt_STWT.csh
adapt_VFS.csh
example.csh
lab_conv.csh

Data:
M001/
M002/
M003/

Data/M001:
M001_B.5mS.cep16
M001_B.5mS.cxt
M001_B.5mS.dcep16
M001_B.5mS.dpow
M001_B.5mS.lpow

Data/M002:
M002_B.5mS.cep16
M002_B.5mS.cxt
M002_B.5mS.dcep16
M002_B.5mS.dpow
M002_B.5mS.lpow

Data/M003:
M003_B.5mS.cep16
M003_B.5mS.cxt
M003_B.5mS.dcep16
M003_B.5mS.dpow
M003_B.5mS.lpow

Etc:
log.tbl
rule

Exe:

Inmatrix:
matrix.200
matrix.300
matrix.400
matrix.50
matrix.600
phone.-
phone.a
phone.b
phone.ch
phone.d
phone.e
phone.g
phone.h
phone.i
phone.j
phone.k
phone.m
phone.n
phone.ng
phone.o
phone.p
phone.q
phone.r
phone.s
phone.sh
phone.t
phone.ts
phone.u
phone.w
```

```
phone.z
phone.zh

Model:
mix12f/
mix12m/
mix12mf/

Model/mix12f:
README
matrix.200
matrix.600

Model/mix12m:
README
matrix.200
matrix.600

Model/mix12mf:
README
matrix.200
matrix.600

Outmatrix:
M001/
M002/
mix2/

Outmatrix/M001:
matrix.200.NEW.HP

Outmatrix/M002:
matrix.200.NEW.HP

Outmatrix/mix2:
matrix.200.NEW.HP
matrix.200.STWT.NEW.HP
matrix.200.STWT.PR.NEW.HP

Src:
Makefile
adapt.c
alloc_fprob.c
alloc_matrix.c
baum_welch.c
cal_norm.c
change_variance.c
cp_mix.c
cut_sp.c
define.h
distortion.c
extern.h
fbcm.c
file_size.c
lab_conv.c
logtbl.c
main_adapt.c
main_train_weight_tied.c
make_list.c
make_std_form_from_ascii.c
mksmsss.c
model_conv.c
print_element.c
read_matrix_ascii.c
read_w_samp.c
set_element.c
set_p_samp.c
set_w_samp.c
smoothing.c
sort_array.c
swap_byte.c
training.c
typedef.h
write_matrix_ascii.c
```

参考文献

[1] 鷹見 淳一, 嵯峨山 茂樹：“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”, 音声研資, SP91-88(1991.12). .c

[2] 鷹見 淳一, 嵯峨山 茂樹：“隠れマルコフ網 (HM-Net) を用いた話者適応”, 音講論, 1-1-8(1992.3).

[3] 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報告, SP92-52, pp. 17–24 (1992.09).