HOMEWORK 3 GMM-based speaker recognition

In this homework, the GMM can be used to recognize the speakers. Please use the speech data in last homework and two more speakers' data given in this homework.

In the previous speaker recognition papers, the training and test sets are (SA1..2,SI1..3, and SX1..3) and (SX4..5). SI(SX)N mean files : SI(SX)Nnn, where 00<nn<00.

The basic procedures, of GMM-based speaker recognition, were : (1) feature extraction, (2) GMM speaker models training, and (3) speaker recognition.

(1) Please use the Tools in HTK to extract the MFCC feature from speech data.
First, you need to download the HTK from http://htk.eng.cam.ac.uk/ Download section.
   (a) You will need to register to HTK site,
   (b) And, you can find and download the HTK manual(v3.3, v3.4)
   (c) Using the following command to find the MFCC feature of speech file,

```
HCopy -C config speaker1_train.pcm fea1_train.mfc
```

Where config is a configure file, in which some parameters in feature extraction were given. The following is an example you can used in this homework.

Filename : conf

```
NATURALREADORDER=TRUE           /* byte order */
NATURALWRITEORDER=TRUE
# Waveform parameters
SOURCEFORMAT=ALIEN              /* file with head length = 0 */
HEADERSIZE=1024
SOURCERATE=625.0                /* sampling rate, unit: 100 nsec */
# Coding parameters
TARGETKIND=MFCC_E               /* output is MFCC and energy */
TARGETRATE=100000.0             /* window shift for analysis, unit: 100 nsec */
SAVECOMPRESSED=F
SAVEWITHCRC=T
WINDOWSIZE=320000.0             /* window width for analysis, unit: 100 nsec */
ZMEANSOURCE=T                   /* remove signal bias */
USEHAMMING=T                    /* take hamming windows before FFT */
PREEMCOEF=0.97                  /* pre-emphasis : 1-0.97Z^-1 */
NUMCHANS=24                     /* number of filter bands */
USEPOWER=F
CEPLIFTER=22                    /* weighting MFCCs */
LOFREQ=0                        /* filter band begin from 0Hz */
HIFREQ=8000                     /* filter band stop at 8000Hz */
NUMCEPS=12                      /* number of cosine transform */
ENORMALISE=T
ALLOWCXTEXP=F
```

Because, the output feature file is binary with 12 bytes file header, please use following command to check the content of feature file. You can see the output feature is 12 order MFCC and energy.

```
HList -C config_hlist -t -o  fea1_train.mfc | more
```

Filename : conf_hlist

```
SOURCEKIND=HTK                /* file type */
SOURCETRATE=100000.0
TARGETKIND=HTK
TARGETRATE=100000.0
```

You change the feature into ACSII in order to used in Matlab code, or read HTK feature directly using codes in https://github.com/ronw/matlab_htk (Sorry, I never try the codes, but it easy if you know the HTK file format.)

(2) The Mixture Gaussian Distribution estimation is the basic skills of speech/speaker recognition. Please find a Matlab code in http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

Please build a GMM model for each speakers, the number of mixtures is 8, 16, 32 and 64, for speakers : FCEG0, MBCG0

(a) The feature vector is high dimensional vector, and difficult to observed. But you can plot the 2-dim histogram (MFCC1-MFCC2) and the distribution you found from GMM and compare their difference.

(b) Put the distribution you found from GMMs from above speakers in order to check their speech characteristics are difference.

(3) Please find the speaker recognition results, i.e., find

$$\underset{i=1,\cdots,N}{Arg\,MAX}\,\log\left(P\left(O_1,O_2,\cdots,O_N \mid \Lambda=(c_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right)\right)$$

where

$$\log\left(P\left(O_1,O_2,\cdots,O_N \mid \Lambda=(c_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right)\right)=\sum_n \log\left(P\left(O_n \mid \Lambda=(c_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right)\right) \quad \forall i=1,\cdots,N$$

(4) Please give the second best speakers in (3).

(5) Please find the speaker recognition results when different lengths of test data were used.

[Hint]

1. If you remove the silence from training and testing data, the result will be better.
2. You can use MFCC only, because using the energy is unfair,
   i.e., TARGETKIND=MFCC in feature extraction.