Modified HTK Toolkit Documentation

To get the latest version from this link:
https://docs.google.com/document/d/1bzV994aXpZ05i9nBb8_noDmq7Zf6YqG8tdHIgJwgsT0/edit?usp=sharing

To train and decode a TVWR system, you need to prepare something else besides the stuff for the standard GMM system:
1. posterior features either in HTK format or sparse format
2. confusion table file may be needed if sparse regression is performed
3. latent variable prior information file may be needed if posterior normalization is needed, which is actually important if the latent variable is associated to a high dimensional context dependent state.
Format of these files can be found in the later section of this document.

HERest options

-u str

r is used to indicate the update of regression parameters, including the static weights

-Y dir ext

dir indicates the posterior file location, it supports two versions of posterior file format: 1). HTK-like binary feature file; 2). Self defined sparse text file. ext indicates the posterior file extension. “-Y dir1 ext1 -Y dir2 ext2 -Y dir3 ext3” may be used to include multiple streams of posterior features.

-dll file

file indicates the confusion table file path. This is primarily used to train TVWR using sparse regression

-upNmodel

It indicates the current objective is to estimate noise model (used in noise adaptive training and noise estimation)

-upAmodel

It indicates the current objective is to estimate acoustic model (used in noise adaptive training)

-xNdir

It enables the location of noise file list changes to the new directory according to -M.

-i num

num indicates the how many EM iterations are used for noise or acoustic model estimation

-nList file

file indicates the path of noise file list

-NAT 20

20 means using head and tail 20 frames to initialize the noise model. This option also enables the VTS adaptation, which works with -nList.

-nt num

num indicates the number of threads to be used for canonical acoustic model estimation.

-W mmf

mmf indicates a MMF file that is used to generate phone posterior feature online

-e mlist

mlist indicates a phone list corresponding to -W.

-tiedStatePos

This enables to use tied state posterior feature, which is specifically implemented for fMPE training.

-J dir

This support loading the fMPE transform.

-h mask

For joint adaptive training, after accumulating the statistics, “-h mask” may be used to search the speaker id given the noise model list. This may be different from the conventional speaker mask, as path and extension are ignored in the noise model list.

HERest Configuration file

HPARMPOS: TARGETKIND = MFCC

This is used to indicate the posterior feature file kind if it is using HTK format. MFCC is just a fake parameter kind, which can be any parameter kind as long as the actual posterior file is the same with this configuration.

POSIZE = 120

This is used to define the dimension of the posterior feature. It is used for regular single stream posterior based TVWR

POSIZE1 = 110

POSIZE2 = 120

POSIZE3 = 130

These are used to indicate the dimension of multiple posterior streams

HMODEL:TEMPCTXSIZE =3 or 1

This is used to enable the temporal context expansion (3) or disable it (1, default)

HMODEL:REGKIND = 0..5

0(NONE_CTX_REG), 1(TEMP_CTX_REG), 2(SPAT_CTX_REG), 3(SPTP_CTX_REG, both spatial and temporal context expansion), 4(HIGH_SPA_REG, sparse regression), 5(TEMP_SPA_REG, temporal expansion of sparse regression). Note that REGKIND should be consistent with TEMPCTXSIZE (either 1 or 3).

HMODEL:CONFUTBLROW =5000

5000 is default row of confusion table. It is not necessary to be exactly the same as the real row. Just make sure it is larger than the real row.

HMODEL:CONFUTBLCOL =5000

Similar to CONFUTBLROW.

FMPE = TRUE |FALSE

This enables fMPE training

HMMIRest options

-dll file

file indicates the confusion table file path. This is primarily used to train TVWR using sparse regression

-Y dir ext

dir indicates the posterior file location, it supports two versions of posterior file format: 1). HTK-like binary feature file; 2). Self defined sparse text file. ext indicates the posterior file extension. “-Y dir1 ext1 -Y dir2 ext2 -Y dir3 ext3” may be used to include multiple streams of posterior features.

-u str

r is used to indicate the update of regression parameters, including the static weights. p is to update the fMPE transform

-W file

file indicates a MMF file that is used to generate phone posterior feature online

-e file

file indicates a phone list corresponding to -W.

-tiedStatePos

This enables to use tied state posterior feature, which is specifically implemented for fMPE training.

-L acc.list

“ls -1 exp/exp_name/hmmc$i/HDR*.acc.[1-3] >acc.list” is used to locate the fMPE statistics

-J dir

This is used to indicate where to load the fMPE transform.

-K dir

This is used to indicate where to store the fMPE transform.

HMMIRest Configuration file

ISMOOTHTAUR=10

I-Smoothing constant term for regression weights.

RE=1

Learning rate for fMPE update

HMMIREST:TRANSKIND=FPROJ

fMPE transformation kind

HMMIREST:TARGETDIM =39

Row number of the fMPE transformation matrix

HMMIREST:INPUTDIMO=39

Dimension of observation

HMMIREST:INPUTDIMP=5769

Dimension of posterior feature, may be the same as a tied-state GMM system used to synthesize the posterior.

HMMIREST:NUMREGIONS=0

Region dependent linear transform, haven’t validated this configuration. So just let it be zero.

HMMIREST:OCTXWIN =1

Region dependent linear transform, haven’t validated this configuration. So just let it be 1.

HMMIREST:PCTXWIN =1

fMPE posterior feature expansion

HMODEL:MINPOS =0.01

A threshold to store sparse high dimensional posterior feature in main memory for fMPE.

HPARMPOS: TARGETKIND = MFCC

This is used to indicate the posterior feature file kind if it is using HTK format. MFCC is just a fake parameter kind, which can be any parameter kind as long as the actual posterior file is the same with this configuration.

HDecode options

-dll file

file indicates the confusion table file path. This is primarily used to load TVWR system based sparse regression

-k 1

TVWR only works with k equal to 1

-Y dir ext

dir indicates the posterior file location, it supports two versions of posterior file format: 1). HTK-like binary feature file; 2). Self defined sparse text file. ext indicates the posterior file extension. “-Y dir1 ext1 -Y dir2 ext2 -Y dir3 ext3” may be used to include multiple streams of posterior features.

-W file

file indicates a MMF file that is used to generate phone posterior feature online

-e file

file indicates a phone list corresponding to -W.

-tiedStatePos

This enables to use tied state posterior feature, which is specifically implemented for fMPE.

-J dir

This is used to indicate where to load the fMPE transform.

-NAT 20

20 means using head and tail 20 frames to initialize the noise model. This option also enables the VTS adaptation. If -nList is specified, VTS will use the noise model from disk.

-nList file

file indicates the path of noise file list

HDecode Configuration file

USEHMODEL = T

It has to be enabled for TVWR decoding

HPARMPOS: TARGETKIND = MFCC

This is used to indicate the posterior feature file kind if it is using HTK format. MFCC is just a fake parameter kind, which can be any parameter kind as long as the actual posterior file is the same with this configuration.

HHEd options

-dll file

file indicates the confusion table file path. This is primarily used to load TVWR model with sparse regression

-r num1 num2

num1 is the expected number of clusters for TVWR parameters. num2 is the max iteration of cluster estimation

-W file

Use MMF file format to store the noise model for offline noise adaptation.

-e mlist

A temporal phone list (i.e. single) for noise model.

-Hp MMF

Add multiple MMFs to obtain a multi-stream GMM system.

-pm file

Build a hybrid CI-NN/HMM system using monophone posteriors. The output file stores the mapping from logical CD states to physical CI state index number (starting from 1).

-pt file

Build a hybrid CD-NN/HMM system using triphone posteriors. The output file stores the mapping from logical CD states to physical CD state index number (starting from 1).

-pl file

Build a hybrid CI-NN/HMM system. The file stores the monophone list, which defines the index information for each monophone.

-h num1 num2

num1 indicates number of streams. num2 indicates the width of each stream. Use this option can reduce the dummy model for hybrid NN/HMM system, such “-h 1 1”.

HHEd Edit file

AP num_1 num_2 prior_file

Add prior information to MMF, num_1 indicates the number of spatial posterior streams, num_2 indicates the number of temporal posterior streams (either 1 or 3). If temporal expansion is enabled, make sure the files with name prior_filel and prior_filer exist. If spatial posteriors are used, prior_file1 prior_file2 prior_file3 … should exist for different stream.

SM

Add multiple MMFs to obtain a multi-stream GMM system.

PC

PMC noise model compensation, works together with noise model from “-W mmf -e single”. Offline adaptation only consider additive noise.

RV

VTS noise model compensation

LV

Extended VTS noise model compensation

HCompV options

-pca num

num indicates how many principal components are retained after PCA.

-fl

This enables converting the posterior to log-posterior

-fc

This enables full covariance estimation

-nf num

num indicates how many utterances there are in total

-nt num

num indicates how many threads are used to perform PCA

-S filelist

filelist can accept loading sparse posterior file list for PCA

-W mmf

-e mlist

These two options load the MMF-like GMM model as posterior synthesizer for PCA

-M dir

This indicates where to store the generated PCA transform

HCopy options

-j file

file indicates the input transform (i.e. PCA or fMLLR) for feature transformation

-fl

This force taking the logarithm of the input feature before output or performing transformation (particularly designed for posterior feature)

-M num num1 num2 ...

This performs the feature merging. num indicates how many streams of features, num1 indicates how width of stream-1 is applied, num2 indicates how width of stream-2 is applied. For this application, script file list would be like: file1 + file2 file3 for each line. file1 and file2 are input, file3 is output.

-c str

m enables CMN, v enables CVN, mv enables CMVN

-g

This converts the global transform to input transform

-Z mmf

-z mlist

These two options load a MMF model to perform fMLLR feature transform or fMPE feature transform.

-fMPE

This indicates the input model is fMPE

-W mmf

-e mlist

These two options load a MMF-like GMM model for synthesizing posterior

-Y dir ext

dir indicates the posterior file location, it supports two versions of posterior file format: 1). HTK-like binary feature file; 2). Self defined sparse text file. ext indicates the posterior file extension.

HCopy Configuration file

HPARM1:TARGETKIND=MFCC
HPARM2:TARGETKIND=MFCC
TARGETKIND=MFCC

Merge two streams of files to one. This is particularly implemented to prepare tandem features. Of course, it can be applied to merge any features.

dnn.map file format: text

Description

List corresponding state index for all logical triphone states, which can be used to convert state-aligned MLF to kaldi format state-label file (used for DNN training).

Tool:

HHEd -A -T 1 -H MMF.txt -w MMF.dnn -pt dnn.map -h 1 1 dummy.hed xwrd.clustered.mlist

MMF.dnn is an self defined hybrid NN/HMM system, which uses the posterior as the acoustic feature for decoding. dummy.hed is an empty edit file.

Example

J-Fr[2] 1682
J-Fr[3] 1689
J-Fr[4] 1767
i-eL+ls[2] 824
i-eL+ls[3] 833
i-eL+ls[4] 352
p-ts+S[2] 397
p-ts+S[3] 408
p-ts+S[4] 414
sil-i+iL[2] 880
sil-i+iL[3] 1105
sil-i+iL[4] 1411

Left is the logical triphone state, right is the physical state index (starting from 1)

Confusion table file format: text

Description:

List most confused latent variables (or states) for each state of GMM system.

Tool:

java DNNConfTableMTfromPDF train.pdf cluster.list pos_dim spa_dir confu_dir/confu.tbl floor njobs

train.pdf is a kaldi format text-ark state-label file used to train NN, which is corresponding to the GMM from TVWR system. cluster.list can be obtained by `grep “~s”

MMF | head -num_states | awk '{print $2}' `. pos_dim is the dimension of the posterior file. spa_dir is the directory of sparse posterior files (yes, it only support sparse posterior file). posterior
Output:

ST_f_2_1:1:55 39:1.33e-01 78:1.18e-01 86:8.59e-02 87:6.19e-02 5:5.94e-02 6:4.18e-02 23:3.26e-02 4:2.54e-02 54:2.53e-02 8:2.12e-02 103:2.09e-02 79:2.00e-02 65:1.74e-02 80:1.58e-02 92:1.57e-02 68:1.51e-02 64:1.47e-02 15:1.41e-02 95:1.31e-02 89:1.20e-02 56:1.18e-02 40:1.13e-02 88:1.04e-02 63:9.47e-03 101:8.69e-03 2:8.50e-03 50:8.16e-03 55:7.72e-03 74:6.77e-03 96:6.37e-03 90:5.97e-03 67:5.89e-03 35:5.82e-03 20:5.52e-03 93:5.30e-03 22:5.25e-03 102:4.77e-03 48:4.28e-03 44:3.81e-03 94:3.64e-03 66:3.37e-03 49:3.09e-03 1:3.03e-03 16:2.65e-03 3:2.62e-03 91:2.62e-03 59:2.10e-03 38:2.04e-03 41:1.90e-03 97:1.89e-03 47:1.80e-03 21:1.80e-03 26:1.65e-03 77:1.41e-03 104:1.13e-03
…
…
…

ST_f_2_1 is the macro symbol used in HTK MMF file, please make sure each tied state must have a macro symbol. :1 is the index of the current physical state. :55 is number of confused latent variables for this state.
The typical setup of HTK has to edit the MMF using HHEd with following configuration:
TI ST_sil_2_1 {sil.state[2]}
TI ST_sil_3_1 {sil.state[3]}
TI ST_sil_4_1 {sil.state[4]}

HTK posterior file format: binary

Description:

Posterior file uses the standard HTK acoustic feature file format. This is primarily used to store CI posterior, as the dimension is relatively low.

Tool:

One example of generating this posterior file is to use following phone recognizer.
http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

Sparse posterior file format: text

Description:

Sparse posterior file is mainly used to store CD posteriors with higher values. Make sure the posterior filename is the same as the acoustic feature file, excluding the extension.

Tool:

nnet-forward --feature-transform=$feature_transform --apply-log=false --use-pos-thresh=1e-5 final.nnet ark:feats.ark ark,t:temp_spa.ark

This is a modified nnet-forward tool with supporting a new option “--use-pos-thresh=1e-5”. Check the Kaldi source code for more details.

java SpaArk2SpaTxt temp_spa.ark cluster.list spa_dir/ prj2mono

SpaArk2SpaTxt is to convert the single file to multiple files for each utterance. cluster.list is optional. prj2mono is either true or false, which is used to indicate whether convert CD posterior to CI posterior. If prj2mono is true, cluster.list must be present; otherwise, “cluster.list” can be any arbitrary string.

Example:

1982 3052
1039:4.18e-03 1409:7.76e-04 2243:9.75e-01 2247:1.72e-02 2404:1.04e-04 2581:1.79e-03 2586:4.80e-04 197:2.22e-04 560:1.57e-04 703:8.06e-04 704:4.39e-04 706:1.45e-04
1039:1.94e-01 1301:7.02e-04 1302:3.42e-04 1347:1.85e-04 1409:2.19e-03 1410:1.30e-04 1432:1.98e-04 1654:5.41e-04 1807:1.08e-04 1996:1.83e-04 2166:5.00e-04 2168:1.78e-04 2193:2.46e-04 2195:1.40e-04 2243:4.79e-01 2247:2.90e-01 2403:1.01e-04 2404:1.42e-02 2479:1.33e-04 2581:5.87e-03 2584:1.38e-04 2586:2.03e-03 2620:5.95e-04 2621:3.48e-04 2686:6.92e-04 2992:1.66e-03
…
…
…

This utterance has 1982 frames, and 3052 is the dimension of this posterior file. Each frame is represented by one line of sparse posteriors elements with format “Idx:Pos”, whose left hand side is the index of this unit (starting from 1 to 3052), right hand side is the posterior value. This text file totally contains 3052+1 lines of strings.

Noise model file list format: text

Description:

A list of noise models are stored for iteratively adaptation and estimation.

Example:

<VECSIZE> 39<DIAGC>
<CLASS>"01ic0201"
<MEAN> 39
-2.172808e+03 -3.138883e+03 -3.837647e+03 -4.200169e+03 -4.053867e+03 -3.462569e+03 -2.675225e+03 -1.963358e+03 -1.333384e+03 -7.685201e+02 -3.014990e+02 -6.220243e+01 -8.388195e+02 0.000000e+00
<MEAN> 39
2.279395e-01 -3.521254e+00 1.950735e-01 2.932994e+00 -2.540134e+00 -2.745022e-01 -1.886705e+00 4.274807e-01 -2.216965e-01 2.682670e-01 6.944975e-01 -5.481966e-01 -5.078110e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00

0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00
<VARIANCE> 39
5.159937e-02 5.762571e-02 1.376679e+01 1.372231e+01 4.544054e-01 1.839877e+00
4.199871e+01 2.739635e-01 2.195181e-01 2.921475e-01 3.487694e-01 5.307188e-01
8.734901e-02 4.014257e-01 6.938969e-02 6.149529e-01 2.216215e-01 1.435288e+00
9.510427e+00 1.530141e-01 7.530764e-02 5.963169e-02 4.158947e+00 2.209365e-01
1.107264e-01 1.012433e+00 4.867991e-02 1.401835e-02 5.138156e-03 6.268457e-03
9.977875e-03 5.857466e-01 2.194352e-02 1.275997e-02 1.446175e-02 2.772568e-01
1.075574e-01 2.223743e-01 1.987818e-01
<CLASS>”01ko0306"
…

<CLASS> is used to indicate the utterance id without extension. Therefore, if joint adaptive training is performed, the speaker mask should not contain the extension.

Latent variable prior information file format: text

Description:

This file is used to insert the prior information in case of TVWR or NN/HMM requiring it.

Tool:

java CompPriorFromPDF train.pdf train.prior num_stat cluster.list

If cluster.list is present, the output is the prior of CI states. Otherwise, the output is the prior of CD states. num_stat is the number of output states.

Example:

120
1:1.25e-02
2:1.58e-02
3:1.17e-02
4:1.79e-02
5:2.25e-02
6:1.49e-02
7:4.79e-02
8:5.20e-02
9:5.16e-02
...

This file contains 120 lines. First line indicates the number of states is 120. Idx:Pos is an representation of state prior probability.

HTK MMF-like additive noise model file for offline adaptation

---

~o

<STREAMINFO> 1 39
<VECSIZE> 39<NULLD><MFCC_D_A_0><DIAGC>
~h "single"
<BEGINHMM>
<NUMSTATES> 3
<STATE> 2
<MEAN> 39
-1.756284e+01 -4.906964e+00 -2.814961e+00 -7.649393e+00 -2.294460e+00
-8.440413e+00 -7.809079e-01 -4.159414e+00 1.415273e+00 -3.874462e+00 3.127136e+00
-3.312915e+00 4.227599e+01 -5.347793e-02 -1.243721e-02 -3.124739e-02 3.826057e-06
2.474860e-02 1.213759e-02 -1.553750e-03 4.678287e-03 9.541146e-03 1.908435e-02
1.911184e-02 7.130384e-03 -4.904363e-02 6.571270e-03 8.163732e-05 3.450272e-03
-1.630215e-03 -3.978393e-03 -4.438540e-03 -1.859494e-03 -6.413622e-04 -1.023481e-03
-4.586778e-03 -1.142294e-03 -1.666660e-03 7.197813e-03
<VARIANCE> 39
2.156805e+01 1.774335e+01 1.801530e+01 2.684799e+01 2.176718e+01 2.454667e+01
2.058136e+01 2.251554e+01 2.810465e+01 2.947150e+01 1.795007e+01 1.512153e+01
2.330469e+01 4.319895e-01 6.410729e-01 7.942138e-01 9.674990e-01 1.224361e+00
1.370376e+00 1.482080e+00 1.652257e+00 1.656083e+00 1.630043e+00 1.502180e+00
1.353418e+00 4.505353e-01 7.415451e-02 1.130399e-01 1.444916e-01 1.813251e-01
2.265475e-01 2.620165e-01 2.837493e-01 3.186830e-01 3.182918e-01 3.124960e-01
2.923114e-01 2.630203e-01 7.411587e-02
<GCONST> 9.134160e+01
<TRANSP> 3
0.000000e+00 1.000000e+00 0.000000e+00
0.000000e+00 6.000000e-01 4.000000e-01
0.000000e+00 0.000000e+00 0.000000e+00
<ENDHMM>

It only stores additive noise model for VTS/PMC/DPMC adaptation. It can be loaded by HHEd -W mmf -e single.list