**Training Flow**

**Files needed**:

1. Corpus = newline separated statements without any punctuations (all chars in lower case)

2. Sound files

3. '.dic' Dictionary file, '.lm' ngram file, and acoustic model files

**Steps**:

**Transform through intermediate tool**:

Corpus + Sound Files -> Transcripts (in the desired format) + Fields file + Renaming and storing of sound files under a single folder

**Training through bw + map\_adapt**:

Transcript + Fields File + Sound Files + Dictionary -> Accoustic model

**Finally, copy:**

Copy accoustic model, dictionary, n-gram files to the respective location in the main project

**Detailed Steps:**

1. **Steps to create lm file and corresconding dictionary**
   1. Create corpus.txt a text file containing all sentences.
      1. The format of corpus.txt file - Each sentence should be in new line and should not have punctuation symbols (, . " " etc).
   2. Goto link [**http://www.speech.cs.cmu.edu/tools/lmtool-new.html**], upload this text file to 'Upload a sentence corpus file:'
   3. Click on COMPILE KNOWLEDGE BASE button, download the obtained files
   4. Obtained '.lm' file and '.dic' contains text in the capital case, convert this text into the small case
2. **Steps to train Acoustic Model**

**[Link: http://cmusphinx.sourceforge.net/wiki/tutorialam]**

[Steps to create sample.transcription and sample.fileids file]

* 1. Copy corpus.txt to 'Training\_Module/CreateTranscription/input\_corpus'
  2. Copy training sound folder corresponding to each trainner to Training\_Module/CreateTranscription/input\_sound\_files **NOTE: Each folder should contain sound files corresponding to all statements in the corpus.txt and the sequence of sound files should be same as the statement in corpus.txt**
  3. Open netbeans project CreateTranscription and run it
  4. Copy '.dic' file obtained in 'Steps: 1.4' to Training\_Module/TrainingAcousticModel and rename it as 'en.dict'
  5. Move all sound files from 'Training\_Module/CreateTranscription/output\_sound\_files' to 'Training\_Module/TrainingAcousticModel'
  6. Move sample.transcription and sample.fileids from 'CreateTranscription/sample\_directory' to TrainingAcousticModel
  7. Goto directory TrainingAcousticModel in terminal and execute following commands to get model

sphinx\_fe -argfile en-model/feat.params \

-samprate 16000 -c sample.fileids -di . -do . \

-ei wav -eo mfc -mswav yes

./bw \

-hmmdir en-model \

-moddeffn en-model/mdef \

-ts2cbfn .cont. \

-feat 1s\_c\_d\_dd \

-lda en-model/feature\_transform\

-cmn current \

-agc none \

-dictfn en.dict \

-ctlfn sample.fileids \

-lsnfn sample.transcription \

-accumdir .

./map\_adapt \

-moddeffn en-model/mdef \

-ts2cbfn .cont. \

-meanfn en-model/means \

-varfn en-model/variances \

-mixwfn en-model/mixture\_weights \

-tmatfn en-model/transition\_matrices \

-accumdir . \

-mapmeanfn en-new/means \

-mapvarfn en-new/variances \

-mapmixwfn en-new/mixture\_weights \

-maptmatfn en-new/transition\_matrices

* 1. Copy 'en-new' folder, 'en.dict' file to 'CoCubesSpeechEvalutionTest/models'
  2. Copy '.lm' file obtain in 'Steps: 1.4' to 'CoCubesSpeechEvalutionTest/models' and rename it to 'eng.lm'

**NOTE: If Sound file are not in 16Khz format then convert them using following steps:**

1. First install ffmpeg command: sudo apt-get install ffmpeg
2. Goto sound directory, then run following command:
3. for f in \*.wav; do ffmpeg -i "$f" -acodec pcm\_s16le -ac 1 -ar 16000 "t${f%.wav}.wav"; done