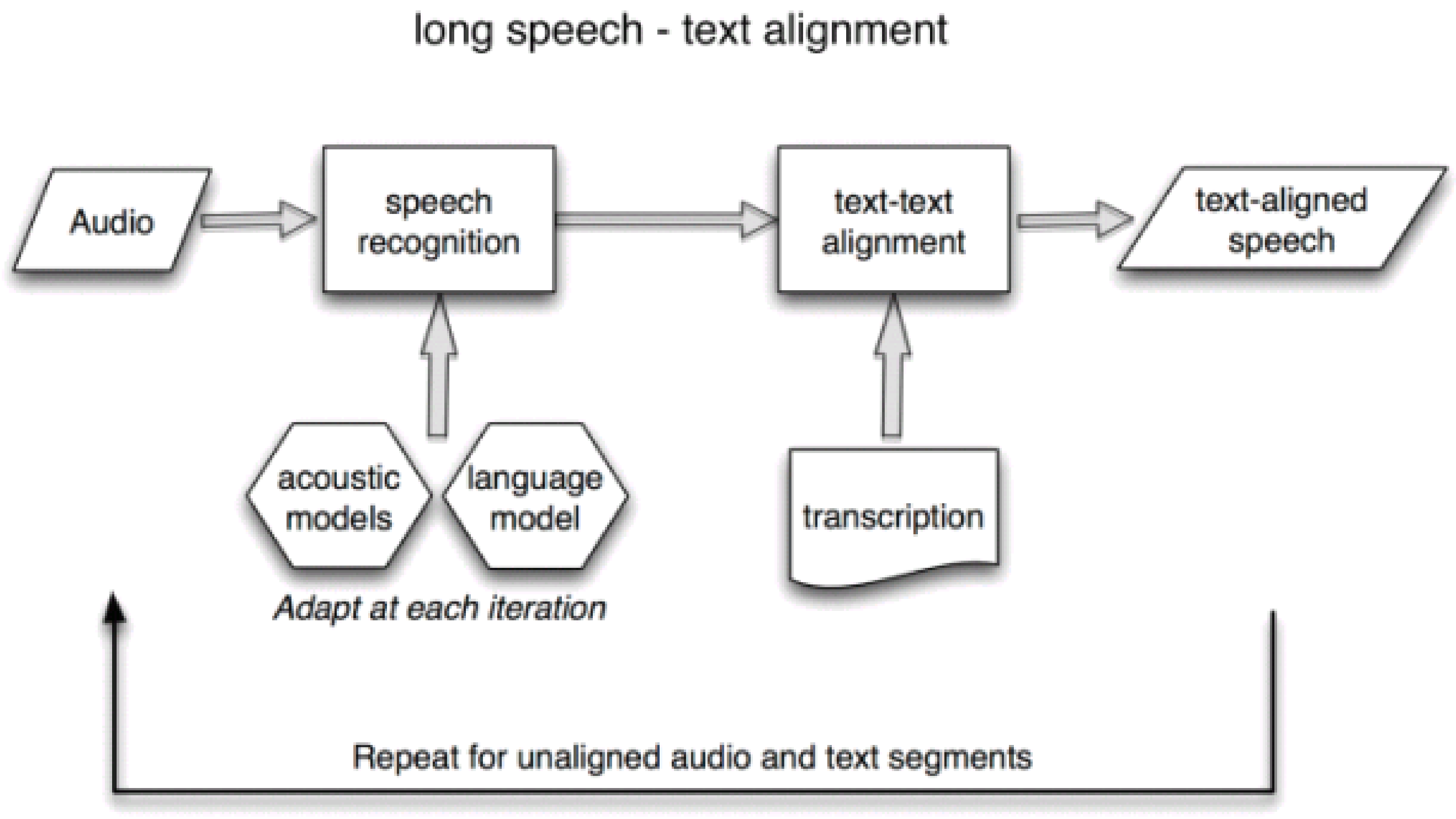
LONG AUDIO ALIGNMENT

Abstract:

- The conventional Viterbi based forced alignment may often be proven inadequate mainly due to mismatched audio and text and/or noisy audio.
- In this project, we work on SailAlign algorithm for long speech-text alignment that circumvents these restrictions which adaptive, iterative speech recognition and text alignment scheme that allows for the processing of very long (and possibly noisy) audio and is robust to transcription errors.

Sail Align:



The Architecture:

- SailTools
  - Library of perl packages, implementing the overall speech-text alignment scheme
  - Has been designed to allow integration of custom speech recognition, acoustic adaptation and language modeling tools
  - Currently SailAlign has only been tested with HTK, SRILM, SCTK.

- Hidden Markov Model Toolkit (HTK)
  - HDecode and HVite are used as speech recognition engines.
  - HERest is used for the adaptation of the acoustic models.
  - HCopy is used for acoustic feature extraction from audio.

- SRI Language Modeling Toolkit (SRILM)
  - ngram and ngram-count are used for language model building.

- SCTK
  - sclite is used for text-text alignment

- Problems faced:

  - The available Sail align toolkit is compatible only with HTK which is old speech recognition engine

  - So, we decided to re-implement the Sail Align algorithm in Kaldi using GMM-HMM

Dataset and Lexicon:

- Training set:
    - Each Mann ki Baath speech module is around 30 minutes in length
    - Around 3 hrs of data has been used for training(6*30 minutes ’data)
- Corpus:
    - The 3 hrs speech data has been divided into around 1 minute samples and corresponding beam width has been taken
- Lexicon:
    - The Lexicon has been built using unified-parser toolkit prepared by DON LABS IITM

Sail Align Algorithm:

```
Require: Audio file and corresponding transcription (word sequence S)
Ensure: Time-aligned transcription (S, T)
 Detect speech regions by Voice Activity Detection (VAD)
 Extract acoustic features A from the audio signal
 E0 ← Generic acoustic models
 U0 ← (A, S) {Unaligned acoustic features and the corresponding word sequence}
 for i=1 to 5 do
      for all N segments in $U_{i-1}$ do
           $A_n$ ←acoustic features of the segment
           $S_n$ ← corresponding word transcript
           Segment $A_n$ in $K_n$ subregions {$A_{n,k}$} of approx-
           imate duration D {Given VAD timestamps, ensure
           that breaks are not within words}
           if  i < 4 then
                Build a trigram language model $L_n$ on $S_n$
         else
               Build a finite state grammar $L_n$ on $S_n$
               if i = 5 then
                       Do not allow insertions or deletions
              end if
        end if
        for k=1 to Kn do
            $(R_{n,k}, T_{n,k}) = \text{SpeechRecognition}(A_{n,k}, E_{i-1}, L_n)$
            {$R_{n,k}$ is the word sequence, $T_{n,k}$ the corre-sponding set of temporal word boundaries}
        end for
    end for
    $(R, F) \leftarrow U_{n,k}(R_{n,k}, T_{n,k})$
    Align word sequences S and R using Dynamic Programming to minimize Levenshtein distance
    $\{(A_{im}, O_{im}, T_{im}), m = 1 \text{ to } M\} \leftarrow$ Subsequences of atleast three aligned words and the corresponding acoustic features {Anchors}
    if i < 4 then
        $E_i \leftarrow$ Adapted acoustic models on$\{(A_{im}, O_{im}, T_{im})\}$ using regression class tree-based Maximum Likelihood Linear Regression (MLLR)
    else
        $E_i \leftarrow E_3$
    end if
    $\{P_j, j = 1 \text{ to } J\} \leftarrow S \backslash U_m O_{im}$
    $\{A_j, j = 1 \text{ to } J\} \leftarrow A \backslash U_m A_{im}$
    $U_i \leftarrow \{(A_j, P_j), j = 1 \text{ to } J\}$ {Collection of unaligned segments and their untimed transcriptions}
end for
$(S, T) \leftarrow U_{i,m}(O_{im}, T_{im})$
```

Implementation:

- The core of the algorithm lies the assumption that the long speech-text alignment problem can be posed as a long text-long text alignment problem given a well performing speech-text conversion tool
- Generic acoustic model
- Initialization
    - mfcc features of test data

  - The audio stream has been segmented into smaller chunks whose duration is constrained by computational limitations of the speech recognition engine used (approximately 10 to 15 seconds in our case). To avoid cutting a word into two, segmentation is guided by a voice activity detection(VAD) module

- Trigram model - To ensure that the speech recognition output will be as close as possible to the reference transcription, a transcription specific language model is built - IRSTLM
- Continuous speech recognition is then applied to identify the lexical content of the individual speech segments
- Using decode obtain the timing of the words
- SCLite - text-text alignment problem can usually be solved quite efficiently even for long text using dynamic programming to minimize the Levenshtein distance between the reference and the hypothesized text

- Acoustic and Language Model Adaptation: To improve noise robustness, we adapt the acoustic models at each iteration in a supervised manner using the reliably aligned regions.

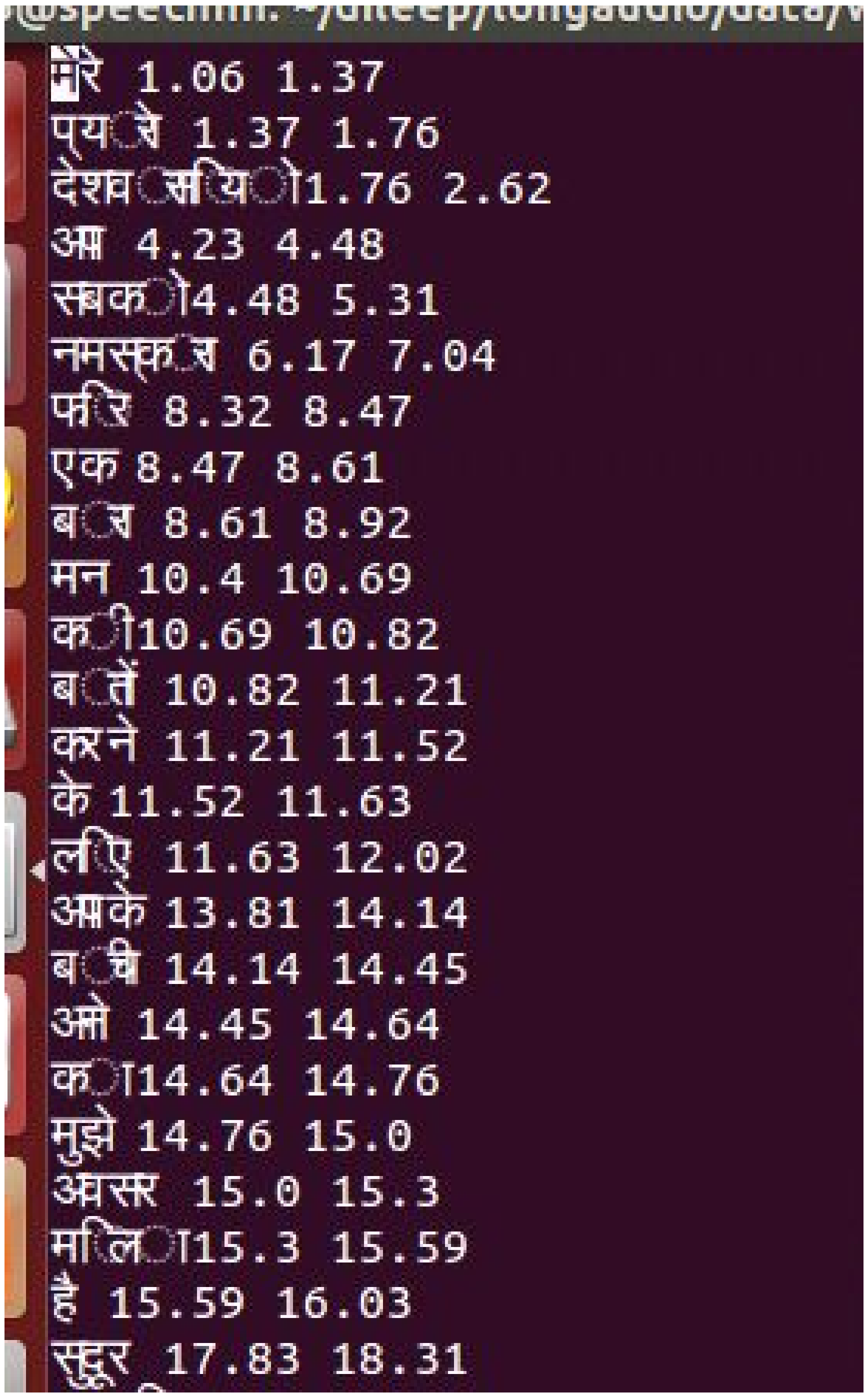
- Maximum Likelihood Linear Regression is applied and adaptation is performed in two steps. First, we train a global transformation and then, for groups of phonemes in which we have sufficient adaptation data, we build a class-based transformation.

- The language models are also updated so that they are trained specifically for each unaligned region.

- This process, i.e., recognition-alignment-adaptation, is iterated three times. In the subsequent two iterations, the acoustic models are not adapted, and the language model is described by a constrained finite state grammar which only allows the expected sequence of words for the segment (and insertions/deletions for the fourth iteration).

- This is expected to further increase the number of aligned regions in the case of very noisy audio.

Results:



```
मेरे 1.06 1.37
प्यारे 1.37 1.76
देशवासियो 1.76 2.62
आज 4.23 4.48
सबको 4.48 5.31
नमस्कार 6.17 7.04
फिर 8.32 8.47
एक 8.47 8.61
बार 8.61 8.92
मन 10.4 10.69
की 10.69 10.82
बातें 10.82 11.21
करने 11.21 11.52
के 11.52 11.63
लिए 11.63 12.02
आपके 13.81 14.14
बीच 14.14 14.45
आने 14.45 14.64
का 14.64 14.76
मुझे 14.76 15.0
अवसर 15.0 15.3
मिला 15.3 15.59
है 15.59 16.03
सुदूर 17.83 18.31
```

References:

- Sailalign - https://github.com/nassosoassos/sail_align/tree/master/docs

- SPEECH RECOGNITION WITH WEIGHTED FINITE-STATE TRANSDUCERS-http://www.cs.nyu.edu/~mohri/pub/hbka.pdf