Building Automatic Speech Recognition (ASR) system using Kaldi toolkit.

Jagbandhu Mishra, Ayush Agarwal, Lalaram Arya

Department of Electrical Engineering
Indian Institute of Technology (IIT) Dharwad

July 19, 2021

Work flow of ASR system

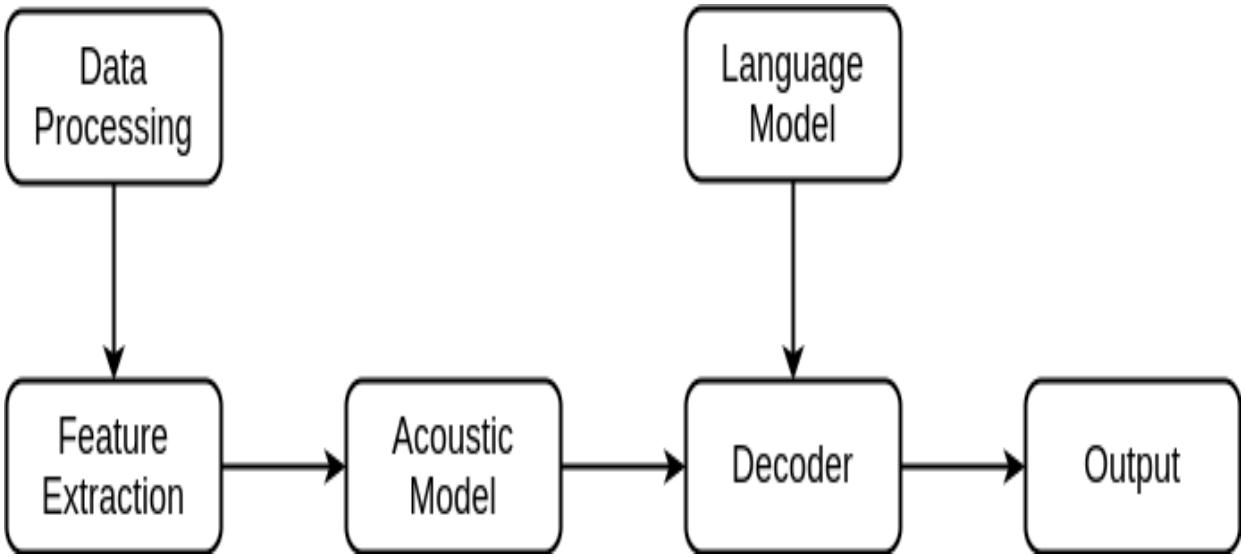

Figure: Flow diagram of ASR

Overview

- Database: Mini-Librispeech
  1. Training: 5 hour
  2. Testing: 2 hour
  3. Language model: small
- Data Download
- Data preparation
- LM preparation: Dictionary and Language model (custom and pre-trained)
- Feature extraction: MFCC-Δ – $\Delta\Delta$
- Mono-phone training, decoding
- Tri-phone training , decoding
- LDA, MLLT and SAT training and decoding

Prerequisite

- cat
- ls
- awk
- grep
- paste
- find

Clone MiniLibrispeech

Clone the miniLibrispeech from github:

- Open the kaldi folder
- Open terminal (using ctrl+alt+t)
- Type cd egs
- Type git clone https://github.com/jagabandhumishra/IEEE-VSSASR-Kaldi-mini-librispeech

Corpus

- Train data
- Test data
- Vocabulary
- Lexicon
- Arpa - small (pruned, 3e-7)

These above files and folders should be in corpus folder:

- To prepare corpus folder type in the terminal:
  - cd IEEE-VSSASR-Kaldi-mini-librispeech/s5
  - mkdir corpus
- Copy these 5 files and folders in corpus folder

Data preparation: speech

- wav.scp $\rightarrow$ utterance - location
- text $\rightarrow$ utterance - text
- utt2spk $\rightarrow$ utterance - speaker
- spk2utt $\rightarrow$ speaker - utterance
- utt2gender $\rightarrow$ utterance - gender

Feature extraction

- Raw speech contains lots of redundant information for ASR task.
- MFCC features can be extracted and used to model the system.
- Kaldi-link,

Language model preparation

- ARPA:
  - Vocabulary
  - Lexicon
  - Arpa: small
- Custom:
  - Vocabulary
  - Lexicon

Dictionary preparation

Create a dictionary (say dict) inside data/local directory:

- extra_questions.txt
- lexicon.txt (word & its phone level break up)
- nonsilence_phones.txt (all the phones excluding silence)
- optional_silence.txt (silence phone)
- silence_phones.txt (silence phone including additional fillers such as bgnoise, chnoise)

Language preparation

A Language directory is created with the below files Kaldi-link :

- L.fst, FST form of lexicon.
- L_disambig.fst, L.fst but including the disambiguation symbols.
- oov.int, mapped integer of out-of-vocabulary words.
- oov.txt, out-of-vocabulary words.
- phones.txt, maps phones with integers.
- topo, the topology of the HMMs we use.
- words.txt, maps words with integers.
- phones/, specifies various things about the phone set.

Mono-phone training

- steps/train_mono.sh
  - To check the model statistics run "gmm-info exp/mono/final.mdl"
  - To see the phone transition run "show-transitions data/lang_nosp/phones.txt exp/mono/final.mdl |less" and then "gmm-copy –binary=false exp_FG/tri_8_2000/final.mdl exp_FG/tri_8_2000/final.txt"

Force alignment

- Phone alignment
  - steps/align_si.sh

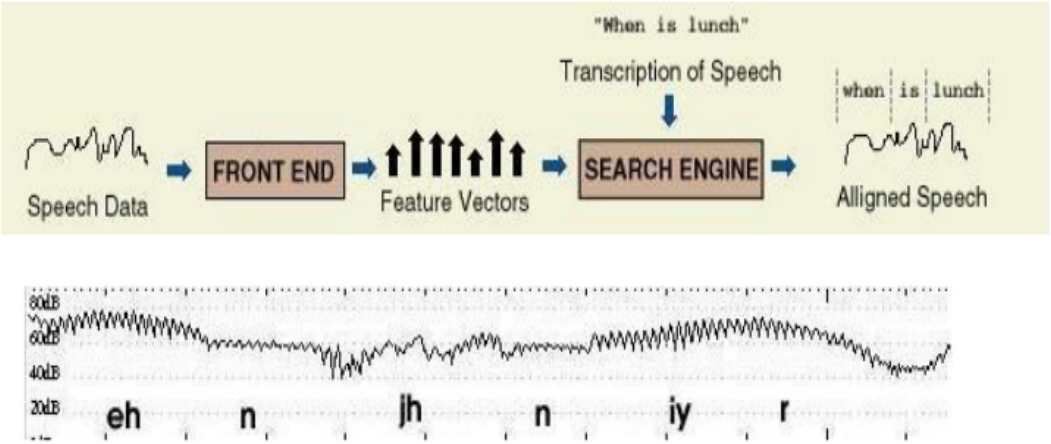

Figure: Phone alignment

- Graph
  - utils/mkgraph.sh
- Decode
  - steps/decode.sh

Tri-phone training

- Mono $\rightarrow$ delta $\rightarrow$ train
  - steps/train_deltas.sh
  - steps/align_si.sh
- LDA (Linear Discriminant Analysis)
- MLLT (Maximum Likelihood Linear Transform)

Performance Evaluation

The Word Error Rate (WER) is a way to measure performance of an ASR.

$$WER = \frac{S + D + I}{N}$$

where,

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions and
- N is the number of words in the reference

Example

Example:

REF: I **** am going to the college

HYP: I can of going to college

Eval I S D

WER = 100 (1+1+1)/6 = 50%

Accuracy= 50%

Acknowledgement

- Prof. Samudravijaya K
- Dr. Sishir Kalita
- Ms. Shruti B.S.
- Speech processing and machine learning group IITDH
- Volunteers

Thank You