Incorporating L2 Phonemes and Utilising L2 Native Speech for Robust L2 English Speech Recognition

(Ref. Interspeech 2023)[Link]

Problem: Building an ASR for L2 English speech

Solution: Extending lexicon with inclusion of L2 phonemes and training ASR with L1 English speech + L2 native speech (in order to show acoustic realisation of L2 phonemes)

Experimental setting:

Dataset type	Dataset	Duration (in Hrs)	Phone set	Train/Test
L1 English	Librispeech-clean-360	360	CMU	Train
L2 Native (Hindi)	IITM Hindi	316	Unified Parser	Train
L1 English	Librispeech-test-clean	5.4	-	Test
L2 English	L2-ARCTIC	3	-	Test

Objectives:

1. Extending lexicon using phoneme transfer rules as IndicTIMIT paper [1].
    a. Creation of common phoneme set
    b. Extending lexicon and converting to common phoneme set
2. Training ASR
    a. Librispeech-clean-360-ASR (Baseline)
    b. Combined_ASR (Proposed)(with English+Hindi+Extended lexion)
3. Testing with Librispeech_clean and L2-ARCTIC-Hindi

1. C. Yarra, R. Aggarwal, A. Rajpal, and P. K. Ghosh, “Indic TIMIT and Indic English lexicon: A speech database of Indian speakers using TIMIT stimuli and a lexicon from their mispronunciations,”in Proc. Oriental COCOSDA. IEEE, 2019, pp. 1–6

Extending lexicon:

1. **CMU**: 40 = 32 (Common) + 8 (Non-common)
2. **Unified Parser:** 66 = 32 (Common) + 34 (Non-Common)
3. **Total:** 74 = 32 (Common) + 8 (CMU) + 34 (Unified Parser)

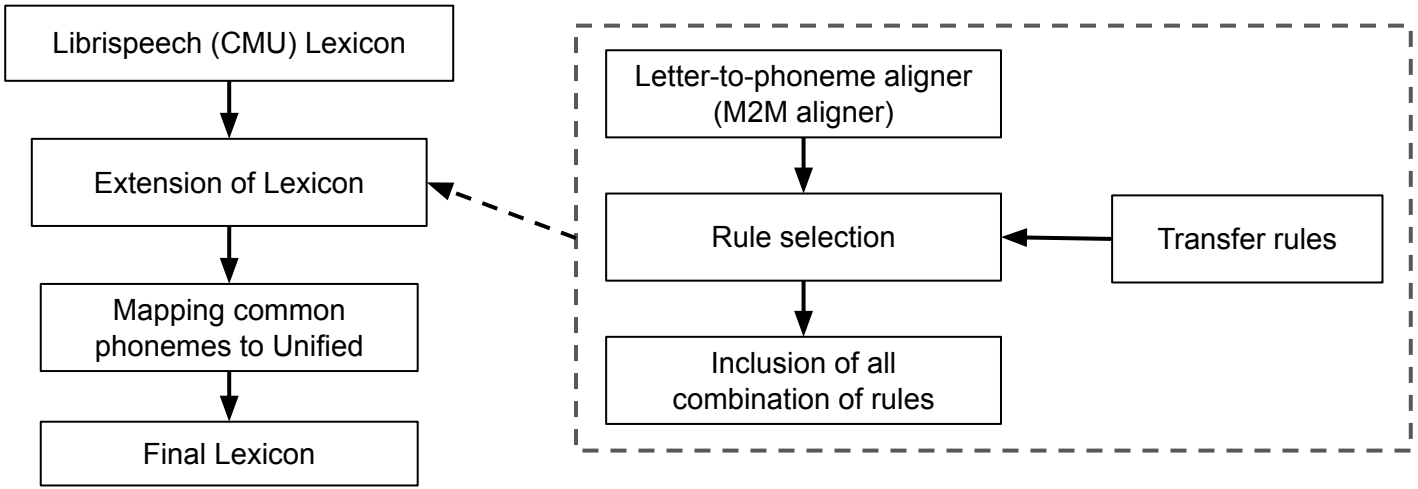
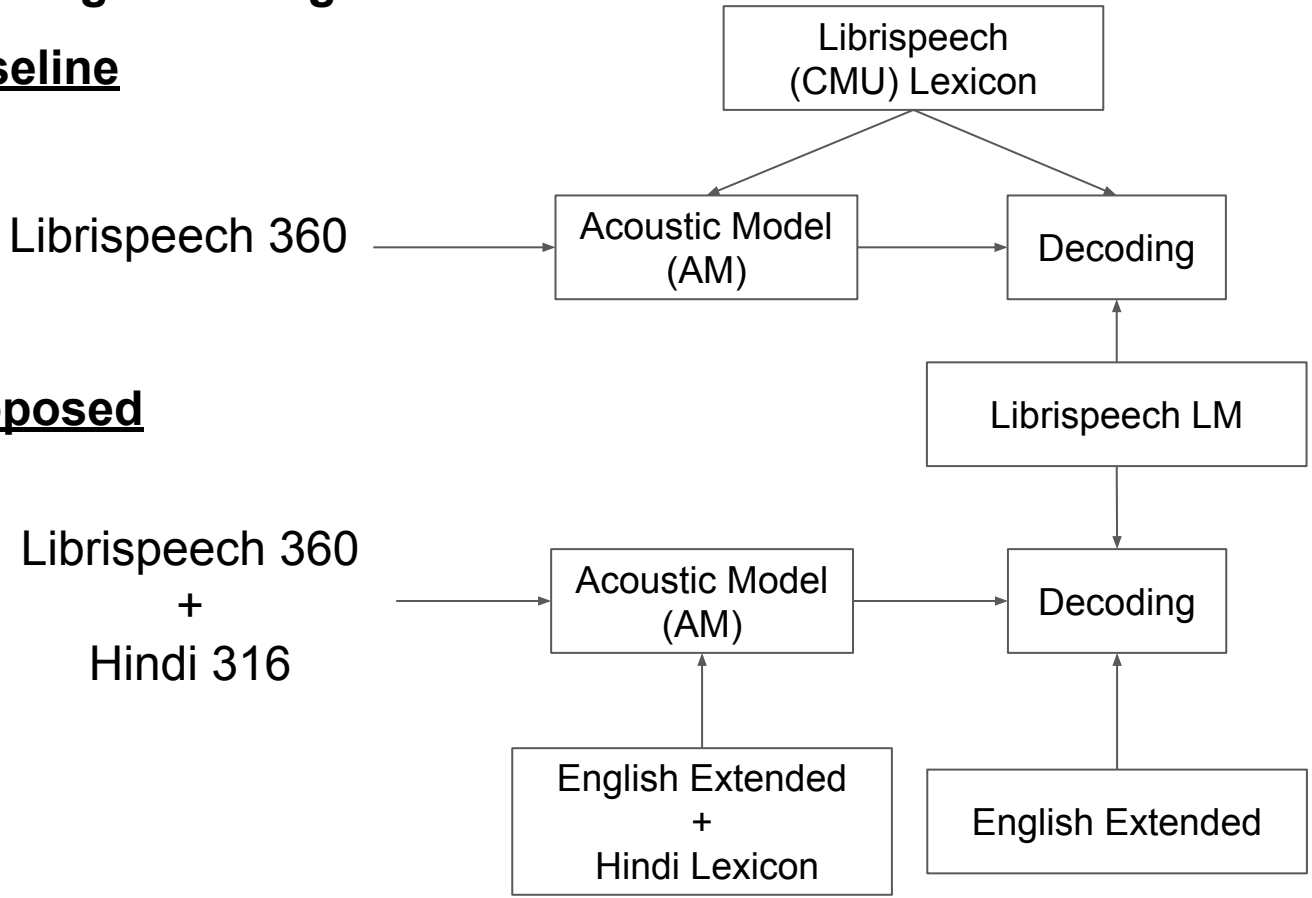


Fig: Flow of extending lexicon

Training & Testing ASR:

Baseline



Proposed

Results:

Table 1: WER in percentages obtained using GMM-HMM Model

Model	Librispeech Test Clean (L1 English)	L2-ARCTIC (L2 English)
Librispeech_360 (Baseline)	16.18	62.83
Combined ASR (Proposed)	20.47 ↑+4.29	56.99 ↓-5.84

Table 2: WER in percentages obtained using DNN-HMM (CNN_TDNN) Model

Model	Librispeech Test Clean (L1 English)	L2-ARCTIC (L2 English)
Librispeech_360 (Baseline)	5.79	35.47
Combined ASR (Proposed)	6.25 ↑+0.46	28.32 ↓-7.15

Combined ASR DNN-HMM model is still training might take 2 to 3 more days (taking longer than expected 60% completed so far started on 25th)

Conclusion:

1. Even though WER on L1-English increased, WER on L2-English reduced.
2. Since our goal is to build better ASR for L2-English speech, the proposed approach can be used for building ASR for L2-English.
3. While Extending lexicon with unified phonemes helped in modelling mispronunciation variations, on the other hand utilising L2 native speech helped in recognising L2 phonemes well.